Exploitation d'une hiérarchie de subsomption par le biais de mesures sémantiques

Emmanuel Blanchard

École Centrale de Nantes **Université de Nantes** **École des Mines de Nantes**

ÉCOLE DOCTORALE S T I M

« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATÉRIAUX »

Année 2008 N° ED 366-???

THÈSE DE DOCTORAT

Spécialité : INFORMATIQUE

présentée et soutenue publiquement par

Emmanuel BLANCHARD

le XX mai 2008
à l'École Polytechnique de l'Université de Nantes

Exploitation d'une hiérarchie de subsomption par le biais de mesures sémantiques

Président :	Jérôme Euzenat, Directeur de Recherches, INRIA Rhône-Alpes
Rapporteurs :	Abdelkader Djamel Zighed, Professeur, Université Lumière Lyon 2
	Giuseppe Berio, Chercheur confirmé, Università degli Studi di Torino
Examinateurs :	Henri BRIAND, Professeur, École Polytechnique de l'Université de Nantes
	Pascale KUNTZ, Professeur, École Polytechnique de l'Université de Nantes
	Mounira HARZALLAH, Maître de conférences, I.U.T. de Nantes

Directeur de thèse : Henri BRIAND
Laboratoire : Laboratoire d'Informatique de Nantes Atlantique (LINA) - UMR 6241 - CNRS
2, rue de la Houssinière – BP 92208 – 44322 Nantes Cedex 3

À ma femme.

Je tiens tout d'abord à remercier ceux qui m'ont initié à la recherche en informatique : Henri Briand, Pascale Kuntz et Mounira Harzallah. J'ai eu la chance de profiter de la longue expérience d'henri pour découvrir le monde de la recherche. Les précieux conseils de Pascale ont largement contribué au bon déroulement de mon travail. Mounira qui a encadré au plus près cette thèse s'est fortement impliquée et je voudrais souligner sa persévérance et sa grande disponibilité à mon égard. Ils ont permis l'aboutissement de cette thèse en stimulant mon goût pour la recherche et en guidant mes réflexions.

Je voudrais remercier les membres du jury pour l'intérêt qu'il porte à mon travail. Je remercie plus particulièrement Djamel Zighed et Giuseppe Berio d'avoir accepté la lourde tâche de relire cette thèse ainsi que Jérôme Euzenat d'avoir accepté de faire partie de mon jury.

Je dirige également ces remerciements vers l'ensemble des collègues de l'IUT et de Polytech pour la chaleur de leur accueil. Merci à tous pour les discussions passionnantes et la bonne humeur générale. Cette thèse a été l'occasion de rencontres très enrichissantes tant d'un point de vue professionnel que personnel. Un grand merci également à ma petite famille !

Table des matières

Liste des figures

Liste des tableaux

Liste des Algorithmes

Introduction

Ingénierie des Connaissances

Le développement de l'informatique s'accompagne d'une évolution des possibilités de stockage (et d'exploitation) des données sous des formes plus ou moins élaborées (fichiers plats, fichiers structurés, bases de données, bases de connaissances). Cet essor a ouvert la voie à l'Ingénierie des Connaissances (IC) qui modélise les connaissances d'un domaine pour les opérationnaliser dans un système destiné à assister une tâche ou le travail intellectuel dans ce domaine (résolution de problème, aide à la décision, consultation documentaire, etc.) [Bac04]. L'Ingénierie des Connaissances est encore une jeune discipline qui, bien que principalement issue de l'intelligence artificielle (IA), entretient des relations de collaboration voire de filiation avec bien d'autres disciplines : linguistique, psychologie, sociologie, logique formelle, etc.

Les sytèmes experts (non supervisés) développés en IA avaient pour objet principal la résolution automatique de problèmes. Les systèmes à base de connaissances développés en IC et ayant succédé aux systèmes experts sont fortement anthropocentrés et visent de nombreuses fonctionnalités :

- le stockage et la consultation de connaissances ;
- le raisonnement automatique sur les connaissances stockées (sans préjugé sur le type de raisonnement à mener) ;
- la modification des connaissances stockées (ajout ou suppression de connaissances) ;
- le partage de connaissances entre systèmes informatiques (problématique d'interopérabilité).

La plupart des systèmes à base de connaissances développés à l'heure actuelle repose sur une masse de connaissances appelée ontologie représentée sous forme d'un réseau de concepts organisés par des relations. La relation principale qui structure les concepts est la relation de subsomption : lorsqu'un concept est plus spécifique qu'un autre on dit qu'il est subsumé par celui-ci (e.g. le concept de *voiture* est subsumé par celui de *véhicule*). La relation de subsomption permet ainsi une hiérarchisation des concepts.

L'exploitation des ontologies repose classiquement sur la logique du premier ordre. Cependant, on a parfois recours à l'évaluation numérique de liaisons entre concepts par le biais de mesures sémantiques.

Mesures sémantiques

La structuration des concepts au sein d'une ontologie définit la sémantique des concepts. Ainsi, les mesures qui exploitent cette structure sont qualifiées de mesures sémantiques. Ces mesures permettent d'évaluer une liaison entre deux concepts d'une même ontologie sur la base des relations qu'ils entretiennent. La signification et le comportement de ces mesures varient selon les besoins applicatifs.

De nombreuses mesures sémantiques ont été développées pour des objectifs applicatifs variés dans des domaines divers (e.g. la linguistique informatique, l'intelligence artificielle, la biologie) aussi bien pour des objectifs académiques qu'industriels. Elles trouvent notamment leur application en recherche d'information pour améliorer la pertinence des résultats et leur classement [LKL93] [Sus93]. Destiné au web sémantique, Corese [CDKFZ04] est un moteur de recherche reposant sur une ontologie qui permet la recherche approximative grâce à une mesure sémantique. Resnik [Res99] a proposé deux applications de sa mesure sémantique au problème de l'ambiguïté du langage naturel. Budanitsky et Hirst [BH01] ont comparé cinq mesures dans un système de détection et de correction de fautes d'orthographe. Dans le domaine de la bio-informatique, Lord et al. [LSBG03] ont utilisé une mesure sémantique pour rechercher une protéine sémantiquement proche d'une protéine donnée.

Quelques mesures [Sus93] [HSO98] [MMRV05] s'emploient à évaluer la proximité sémantique de deux concepts en utilisant plusieurs relations entre les concepts. Cependant, la majorité des propositions se restreint à la relation de subsomption. Il s'agit tout d'abord d'une relation fortement structurante [RMBB89] et commune à toutes les ontologies. L'organisation hiérarchique induite par la relation de subsomption permet une factorisation des caractéristiques communes au niveau des subsumants communs. On dit alors que chaque concept hérite des caractéristiques de ses subsumants. Ce mécanisme d'héritage permet donc de réduire grandement la nécessité de spécifier des caractéristiques redondantes. L'héritage est une propriété fondamentale de la relation de subsomption qui la distingue des autres relations. La relation de subsomption est une relation transversale à toute les autres.

Si on s'attache à la signification des mesures sémantiques, on peut distinguer deux types de significations distincts [Res99] : degré de ressemblance et force du lien. La force du lien est une notion plus large que le degré de ressemblance : deux concepts qui se ressemblent sont nécessairement liés sémantiquement (e.g. voiture et vélo) tandis que deux concepts qui sont fortement liés sémantiquement ne se ressemblent par forcément (e.g. voiture et carburant). De façon duale, il s'agit parfois de l'absence de ressemblance ou de lien qui est évaluée. Notons également que la signification d'une mesure est liée aux propriétés mathématiques qu'elle respecte (e.g. maximalité, inégalité triangulaire).

Bien que les contributions soient très diverses et souvent indépendantes les unes des autres, deux approches bien distinctes émergent de la littérature :

- La hiérarchie est considérée comme un graphe. Dans le cas d'un arbre, la notion de profondeur est souvent utilisée. De manière plus générale, le plus court chemin entre deux concepts renseigne sur leur proximité sémantique.
- On considère un concept selon la quantité d'information qu'il renferme.

On parle de contenu informationnel $\psi(c_i) = -\log \mathrm{P}(c_i)$ d'un concept c_i où $\mathrm{P}(c_i)$ est la probabilité pour une instance quelconque d'appartenir au concept c_i [Res93]. On peut alors distinguer la quantité d'information commune à deux concepts et celle qui les différencie.

La plupart des mesures sémantiques que l'on peut répertorier sont définies de manière *ad hoc* ce qui rend plus difficile leur comparaison et leur réutilisation. D'autre part, leur signification (ce qu'elles évaluent concrètement) est variable et rarement formellement explicitée. Dans ses travaux, Lin [Lin98] initie une étude théorique de la notion de similarité qui fait le lien avec les mesures sémantiques. Néanmoins, il manque un cadre théorique général pour supporter la comparaison et l'étude des mesures sémantiques dans leur ensemble.

Unified Enterprise Modelling Language (UEML)

Le réseau d'excellence INTEROP-NoE[1] (*Interoperability Research for Networked Enterprise Applications and Software*), est un projet qui s'est déroulé sur 42 mois (Novembre 2003 - Avril 2007) et coordonné par l'université de Bordeaux 1 avec 47 partenaires et plus de 300 chercheurs. C'est dans le cadre de ce projet qu'a émergé UEML qui cherche à faire face au problème de la multiplicité des langages de modélisation d'entreprise.

Pour jouer un rôle important ou au moins survivre dans un monde économique en évolution permanente, les entreprises doivent avoir une vision claire de leur propre structure. Elles ont recours à la modélisation d'entreprise (*Enterprise Modelling*) qui est l'ensemble des activités et processus utilisés pour développer les diverses parties d'un modèle d'entreprise [PD02]. Les langages de modélisation d'entreprise (*Enterprise Modelling Languages*) permettent le développement de tels modèles d'entreprise. Un langage de modélisation d'entreprise définit les constructs génériques du modèle pour une modélisation d'entreprise adaptée aux besoins des gens qui créent et utilisent le modèle d'entreprise. Selon [Ver02], le nombre conséquent de langages de modélisation d'entreprise existants crée une situation difficile pour les utilisateurs désirant utiliser la modélisation d'entreprise (beaucoup de langages de modélisation, vocabulaire et paradigmes instables, incompatibilité des outils de modélisation, peu de fondements formels).

L'objectif de UEML est de supporter l'utilisation intégrée des modèles d'entreprises définis dans des langages différents. UEML est conçue comme un mécanisme pour interconnecter des langages différents et leurs modèles.

Contributions de la thèse

Les contributions de la thèse participent à l'étude des mesures sémantiques pour l'Ingénierie des Connaissances tant d'un point de vue pratique que théorique. Nous proposons un cadre formel centré autour de la notion de contenu informationnel et faisons l'analogie avec des mesures usuelles de la littérature. Le développement d'un applicatif (basé sur l'outil Protégé) implémente notre

[1] http ://www.interop-noe.org

approche et fournit ainsi un environnement d'étude des mesures sémantiques sur une hiérarchie de subsomption réelle.

Notion de contenu informationnel

Nous mettons en évidence le rôle du contenu informationnel introduit par Resnik [Res93] en proposant diverses approximations $\widehat{\mathrm{P}}$ de la mesure de probabilité P sur laquelle il repose. Ces approximations relèvent de trois approches : (1) ascendante, (2) descendante et (3) mixte qui permettent de considérer divers aspects de la hiérarchie de subsomption qui contribuent à qualifier la proximité sémantique des concepts. Nous élargissons cette notion pour le calcul du contenu informationnel global $\psi^{\cup}$ d'un sous-ensemble de concepts ainsi que le contenu informationnel qu'ils partagent $\psi^{\cap}$. Cela nous permet de traiter le cas de l'héritage multiple et l'évaluation de la ressemblance entre sous-ensembles de concepts.

Analogie avec des mesures usuelles

Nous montrons que les mesures sémantiques qui considèrent la structure hiérarchique comme un graphe peuvent être exprimées à l'aide du contenu informationnel en choisissant l'approximation appropriée. Le contenu informationnel est donc une notion clef pour la définition des mesures sémantiques. Nous dressons une typologie des mesures définies sur une représentation ensembliste de manière à mettre en évidence le fait que les mesures sémantiques respectent des schémas bien connus (e.g. Jaccard [Jac01], Dice [Dic45]). Par analogie, nous proposons un cadre général pour l'étude et la définition de mesures sémantiques qui fédère les principaux travaux sur les mesures sémantiques. Cette analogie singulière qui repose sur la notion de contenu informationnel ouvre également la voie pour la définition de nouvelles mesures et notamment des mesures asymétriques. En définitive, nous montrons que la définition d'une mesure sémantique peut se résumer au choix d'une mesure ensembliste et d'une approximation de la probabilité nécessaire au calcul du contenu informationnel.

Plug-in SymanticTab pour Protégé

SymanticTab est un outil dont l'objectif est de permettre d'analyser le comportement des similarités sémantiques sur une hiérarchie de subsomption réelle. Il s'agit plus spécifiquement d'un plug-in pour l'éditeur d'ontologies Protégé[2] qui implémente notre approche afin d'aider un utilisateur dans le choix d'une similarité sémantique. Ce plug-in a également été adapté pour permettre l'étude de mesures entre deux constructs (Les constructs sont des briques de base d'un langage de modélisation d'entreprise que l'on est amené à considérer comme des sous-ensembles de concepts).

Notre travail a été guidé par deux objectifs applicatifs : (1) fournir un outil (SymanticTab) d'aide au choix d'une similarité sémantique adaptée aux besoins d'un utilisateur et (2) développer un outil (*UEMLBase Correspondance*

[2] http ://protege.stanford.edu/

Analyser) pour l'analyse de correspondances entre constructs UEML. Pour le développement de SymanticTab, nous nous sommes limités à la définition de mesures de similarité entre deux concepts. L'analyse de correspondance entre constructs a nécessité l'extension de nos travaux à des mesures asymétriques et ce entre deux sous-ensembles de concepts.

Organisation de la thèse

Cette thèse se décompose en trois parties. La première partie constituée de deux chapitres s'intéresse aux mesures sémantiques comme outil pour la gestion des connaissances. Dans le chapitre 1, nous définissons la notion d'ontologie avant de présenter la terminologie et les structures mathématiques ayant trait à la hiérarchie de subsomption. Le chapitre 2 constitue un état de l'art sur les mesures sémantiques dans lequel nous reprenons quelques applications qui soulignent leur utilisation multi-disciplinaire. Nous présentons ensuite les principales mesures sémantiques de la littérature.

La seconde partie découpée en quatre chapitres est consacrée à la proposition d'un cadre théorique pour l'analyse, la comparaison et la définition de mesures sémantiques. Le chapitre 3 reprend de nombreux travaux adaptés à une représentation ensembliste des concepts. Dans le chapitre 4, nous exposons le principe de l'interpétation extensionnelle d'un arbre de subsomption. Nous reprenons ensuite la notion clef de contenu informationnel en proposant des approximations de la mesure de probabilité sur laquelle il repose. Dans le chapitre 5, nous proposons un cadre fédérateur par le biais d'une analogie qui nous permet de réécrire les mesures sémantiques présentées au chapitre 2 grâce au contenu informationnel. Dans le chapitre 6, nous élargissons notre proposition à l'exploitation d'une hiérarchie de subsomption acceptant l'héritage multiple.

La troisième partie traite de l'évaluation des mesures sémantiques. Le chapitre 7 fournit des éléments statistiques basés sur le réseau sémantique WordNet. Le chapitre 8 est consacré à la présentation du plug-in Protégé qui implémente notre approche.

Première partie

Mesures sémantiques en gestion des connaissances

1 Contexte

Sommaire

Résumé

Ce chapitre a pour objectif de situer les contributions de cette thèse au sein de l'Ingénierie des Connaissances. Il pose les prérequis nécessaires à la lecture de ce document. Dans un premier temps, nous proposons une synthèse sur la notion d'ontologie traitant des aspects théoriques jusqu'à la mise en oeuvre. La seconde partie est dédiée exclusivement à l'étude de la hiérarchie de subsomption avec en préambule l'explicitation de la terminologie associée qui est par ailleurs employée tout au long de ce manuscrit. Nous présentons ensuite les structures mathématiques qui nous servent de support à la modélisation d'une hiérarchie de subsomption ainsi que dans des cas plus restrictifs où l'on parle de treillis et d'arbre. Enfin, nous évoquons brièvement le principe de construction d'un treillis de galois qui met en lumière la relation implicite entre extension et intension d'un concept dans une hiérarchie de subsomption.

1.1 Introduction

La volonté de concevoir des systèmes capables de reproduire un comportement proche de celui de l'être humain dans ses activités de raisonnement se décline selon deux courants de pensée résumés dans la phrase : « making a mind versus modelling the brain ».

La première approche (« making a mind ») a conduit à l'intelligence artificielle (IA) actuelle et aux systèmes à base de connaissances raisonnant sur des données symboliques. La seconde (« modelling the brain ») a débuté avec les travaux en reconnaissance des formes. Ces recherches ont abouti aux modèles connexionnistes comme les réseaux de neurones. Ils reposent sur les principes suivants :

- la délocalisation des connaissances à travers la structure du système et dans les connexions entre éléments, ajustées au cours d'un apprentissage préalable ;
- le traitement réparti de l'information (avec des mécanismes de propagation associés), par opposition au raisonnement symbolique explicite : c'est de la complexité du réseau, et en particulier du très grand nombre de processeurs élémentaires, qu'émergent les comportements « intelligents » de tels modèles.

Les modèles connexionnistes, sans représentation explicite des connaissances, n'ont en général pas la capacité d'expliquer leur propre fonctionnement, c'est à dire les raisonnements qu'ils mènent. Le raisonnement sur des données symboliques nécessite leur formalisation selon un certain mode de représentation, qui peut être défini comme un ensemble de méthodes de codage symbolique. Un mode de représentation associe ainsi deux aspects imbriqués, même parfois confondus :

- une structure de données permettant de représenter la connaissance à coder ;
- une ou plusieurs méthodes d'exploitation de cette connaissance permettant, par un raisonnement, de produire de nouvelles connaissances.

Les premiers systèmes « intelligents » développés en IA et appelés systèmes experts ont permis des avancées dans le domaine de la modélisation des connaissances mais restent limités vis-à-vis des objectifs ambitieux du programme initial. L'objectif de ces systèmes était de stocker les connaissances d'un expert afin de reproduire un comportement similaire. Pour répondre à cette attente, des outils non supervisés ont vu le jour. La non-intervention de l'expert dans le processus est confrontée à la réticence de celui-ci à livrer ses connaissances d'expertises. Aussi, le fait qu'il n'intervienne pas induit un stockage exhaustif de toutes les connaissances nécessaires aux raisonnements, ce qui semble ambitieux. De plus, ces systèmes qui n'expliquaient pas ou pas suffisamment leurs raisonnements apparaissaient donc à l'utilisateur comme des « boîtes noires ». Les résultats obtenus par de tels systèmes étaient difficilement vérifiables et perdaient ainsi beaucoup de leur valeur.

Si les systèmes experts n'avaient à l'origine pour objet principal que la résolution automatique de problèmes, les systèmes à base de connaissances qui leur ont succédé sont en revanche des sytèmes supervisés qui visent l'intégration de plusieurs fonctionnalités : le stockage et la consultation des connaissances, le rai-

sonnement sur les connaissances stockées (sans préjugé sur le type de raisonnement à mener), la modification des connaissances stockées (ajout ou suppression de connaissances), et le partage de connaissances entre systèmes informatiques (problématique d'interopérabilité). Les ontologies jouent un rôle central dans la plupart des systèmes à base de connaissances développés à l'heure actuelle en Ingénierie des Connaissances (IC). Les ontologies sont issues des réseaux sémantiques liés aux travaux de Quillian [Qui68] et Collins [CL75] sur la mémoire sémantique humaine.

1.2 Les ontologies

Le terme « ontologie[1] » a été emprunté à la philosophie pour les besoins de la gestion des connaissances par des systèmes informatiques. Les ontologies suscitent un grand intérêt dans la communauté scientifique du fait des possibilités qu'elles offrent. En effet, les ontologies visent le développement d'artefacts informatiques pour une véritable gestion des connaissances assistée par ordinateur.

1.2.1 Tentative de définition

Comme l'explicite Gruber [Gru93], une masse de connaissances représentée formellement est basée sur une conceptualisation qui selon Genesereth et Nilsson [GN87] correspond aux objets, concepts, et autres entités supposées exister dans un certain domaine d'intérêt et les relations qui les organisent. Par ailleurs, les spécialistes des ontologies s'accordent pour considérer que les primitives cognitives de base d'une ontologie sont les concepts et les relations entre ces concepts. La construction d'une ontologie intervient donc après qu'un travail de conceptualisation ait été mené à bien pour organiser les concepts dans un réseau de relations.

Gruber ajoute que la conceptualisation est une vue simplifiée abstraite du monde que l'on veut représenter dans un but donné. Cette précision souligne le fait qu'il y a potentiellement une multitude de conceptualisations différentes d'un même domaine suivant l'angle de vue qui amène à l'obtention d'une vue simplifiée et l'objectif qui guide implicitement la conceptualisation.

Selon la définition de Gruber [Gru93] qui fait référence en Ingénierie des Connaissances, une ontologie est une spécification explicite d'une conceptualisation. Studer et al. [SBF98] affine la définition de Gruber en considérant les travaux de Borst [Bor97] pour aboutir à la définition suivante : une ontologie est une spécification explicite formelle d'une conceptualisation partagée. Le caractère explicite d'une ontologie signifie que les primitives de base que sont les concepts et les relations sont explicitement définies. Le terme formel impose que l'ontologie soit manipulable au sein d'un système informatique ce qui exclut l'utilisation du langage naturel. Une ontologie est issue d'une conceptualisation partagée, c'est-à-dire qu'elle rend compte d'une connaissance consensuelle, qui ne reflète pas la pensée d'un seul individu mais qui au contraire est acceptée par une certaine communauté.

[1] Selon Welty et Guarino [WG01], ce terme désigne à l'origine une discipline de la philosophie qui étudie ce qui existe et la nature des choses.

Nous considérons également la définition de Guarino et Giaretta [GG95] qui complète celle de Gruber en précisant qu'une ontologie est une spécification partielle d'une conceptualisation en ce sens qu'une conceptualisation ne peut pas toujours être entièrement formalisée, du fait d'ambiguïtés ou du fait qu'aucune représentation de leur sémantique n'existe dans le langage de représentation choisi.

Le processus de construction de l'ontologie a fait l'objet d'études ayant abouti à des méthodologies couvrant différentes parties de ce processus. Malgré la proposition de divers critères pour supporter le processus de construction d'une ontologie, aucune méthodologie générale ne s'est encore imposée à l'heure actuelle. Toutefois, la plus célèbre d'entre elles est la méthodologie Methontology [FGPJ97] qui distingue dix étapes :

1. construire le glossaire des termes qui seront inclus dans l'ontologie, préciser leur définition en langage naturel, identifier leurs synonymes et leurs acronymes ;
2. construire des taxinomies de concepts pour les classifier ;
3. construire des diagrammes de relations binaires *ad hoc* pour identifier des relations *ad hoc* entre les concepts d'une même ontologie et également entre les concepts d'ontologies différentes ;
4. construire le dictionnaire de concepts qui inclut, pour chaque concept, ses attributs d'instance, ses attributs de classe et ses relations *ad hoc* ;
5. décrire en détail chaque relation binaire *ad hoc* qui apparaît dans le diagramme de relations binaires *ad hoc* et dans le dictionnaire de concepts ;
6. décrire en détail chaque attribut d'instance qui apparaît dans le dictionnaire de concepts ;
7. décrire en détail chaque attribut de classe qui apparaît dans le dictionnaire de concepts ;
8. décrire en détail chaque constante (les constantes donnent des informations sur le domaine de connaissances) ;
9. décrire les axiomes formels ;
10. décrire les règles utilisées pour contraindre le contrôle et pour inférer des valeurs aux attributs.

1.2.2 Mise en oeuvre

La mise en oeuvre des ontologies nécessite le choix d'un langage de représentation. Celui-ci doit permettre un traitement automatique en machine tout en restant compréhensible par l'humain de manière à permettre une réelle interaction entre le système et son utilisateur. Il doit également garantir la portabilité de l'ontologie dans un objectif de partage et de réutilisation.

Manipulable en machine et compréhensible par l'utilisateur

Les ontologies sont non seulement destinées à être manipulées de manière automatique par des systèmes informatiques mais doivent également permettre un

dialogue, une coopération entre le système et l'utilisateur humain. Tout d'abord, il est impératif d'utiliser un langage informatique pour que la machine soit en mesure de traiter les connaissances de l'ontologie. Cependant, la compréhension de l'utilisateur nécessite une certaine continuité avec le langage naturel. Ceci est mis en oeuvre grâce au méta-langage de balisage XML[2]. XML reste cependant un langage informatique donc plus ou moins facilement appréhendable pour un non-informaticien. Mais du fait que XML soit un standard du W3C[3] libre de droit, de nombreux outils (e.g. Protégé, Kaon) offrant une interface de visualisation ont vu le jour ; ils facilitent ainsi l'interaction homme-machine.

Pour permettre le raisonnement, le langage utilisé doit être doté d'une sémantique formelle (c'est-à-dire qui a une équivalence en logique du premier ordre). Le langage OWL[4] fondé sur RDF(S)[5] spécifie une syntaxe XML et permet véritablement de représenter une ontologie. Parce qu'une plus grande expressivité entraîne une plus grande complexité, OWL fournit trois sous-langages d'expressivité croissante (OWL Lite, OWL DL et OWL Full). Ces langages participent au développement du Web Sémantique initié par Berners Lee [BLHL01].

Portabilité

Le recours à l'informatique permet la duplication des données et leur diffusion sur le web. On note deux aspects qui participent à la portabilité d'une ontologie : l'émergence du standard OWL pour la représentation des ontologies mais aussi le découplage entre sémantique formelle et opérationnelle.

OWL s'est rapidemment imposé comme le langage standard pour la représentation des ontologies. L'existence d'un tel standard est indispensable pour faciliter leur portabilité. Tous les outils logiciels qui désirent contribuer au développement et à l'exploitation des ontologies prennent en charge ce langage qui constitue actuellement un format d'échange standard.

Au découplage entre représentation des connaissances et mécanismes inférentiels, déjà existant dans les systèmes experts, s'ajoute le découplage entre la sémantique formelle d'un domaine, qui ne fait que contraindre l'interprétation des connaissances, et la sémantique opérationnelle qui précise la façon dont ces connaissances vont être mises en oeuvre dans le SBC pour raisonner [Für04]. En d'autres termes, lors de la construction d'une ontologie, on ne préjuge pas de la façon dont seront utilisées les connaissances pour raisonner. Ceci participe à la portabilité des ontologies nécessaire au partage et à la réutilisation des connaissances. Le langage OWL permet le raisonnement grâce à sa sémantique formelle mais n'offre pas de mécanisme d'inférence qui en ferait un langage opérationnel.

Selon Guarino [Gua98], une ontologie décrit au moins une hiérarchie de concepts liés par la relation de subsomption à laquelle s'ajoute parfois des axiomes pour exprimer d'autres relations entre les concepts et pour contraindre leur interprétation. Dans le paragraphe suivant nous nous attardons sur la notion de hiérarchie de subsomption qui est au coeur de nos travaux.

[2]http ://www.w3.org/XML/
[3]World Wide Web Consortium
[4]http ://www.w3.org/2004/OWL/
[5]http ://www.w3.org/RDF/ et http ://www.w3.org/TR/rdf-schema/

1.3 La hiérarchie de subsomption

De nombreux modèles actuels tels que ceux basés sur les logiques de description [Neb90] [BMNPS91] [Nap97] et les graphes conceptuels [Sow84] [eMM92], structurent les connaissances autour d'une hiérarchie de subsomption. Les principes sous-jacents à ces modèles sont issus des travaux sur les réseaux sémantiques [Qui68] et les frames [Min75].

Les réseaux sémantiques mis en oeuvre par Quillian [Qui68] ont été conçus à l'origine comme un modèle psychologique explicite de la mémoire associative humaine. Ils sont fondés sur un modèle de graphe permettant de combiner la représentation des concepts par l'intermédiaire des noeuds et des relations entre concepts par des arcs orientés. Les étiquettes sur les arcs spécifient le type de la relation entre deux concepts. Ces relations correspondent par exemple à des liens de causalité, des relations spatiales ou temporelles ou encore des relations de spécialisation ou de composition entre concepts.

En 1975, Minsky [Min75] introduit un formalisme de représentation des connaissances centré sur la notion de « frame ». Le principe du modèle des frames est de décomposer les connaissances en classes (« frames ») qui représentent les concepts du domaine. À une frame est rattaché un certain nombre d'attributs (« slots »), chaque attribut pouvant prendre ses valeurs parmi un ensemble de facettes (« facets ») [KLW95]. Les classes ainsi définies sont structurées selon la relation de spécialisation.

Dans la suite de ce chapitre, nous définissons ce qu'est une hiérarchie de subsomption. Nous précisons la terminologie ainsi que les structures mathématiques associées à cette notion. Les notations introduites au cours de cette présentation sont reprises dans la suite de ce manuscrit.

1.3.1 Terminologie

Concept

La notion de concept peut être appréhendée à travers la définition de la fonction caractéristique d'un *concept* selon Bournaud [Bou96] qui s'inspire des écrits de Frege [Fre71].

Définition 1.1 *[Bou96] La* fonction caractéristique d'un concept *est une fonction définie sur un domaine de référence, qui prend ses valeurs dans le domaine {vrai, faux}. Cette fonction discrimine les individus auxquels s'applique le concept –la fonction prend la valeur vraie et les individus, appelés* instances *du concept, sont dit recouverts par le concept– de ceux auxquels il ne s'applique pas.* L'extension *d'un concept est l'ensemble des instances de ce concept.*

Il y a trois façons de faire référence à un concept [Gro02] :

Par son contenu (l'être). L'être est l'extension du concept. C'est l'ensemble des instances du concept (les choses existantes auxquelles le concept s'applique) ;

Par sa définition (l'essence). L'essence est la condition d'appartenance à la

classe. On donne un prédicat ou une définition (une condition) qui permet de créer le concept en intension ;

Par son nom (terme univoque qui abrège la définition). Le nom du concept est un abrégé ultime de la définition. Néanmoins, le nom est avant tout une commodité, un code de reconnaissance, qui est difficilement utilisable si l'on fait abstraction de sa définition complète (ambiguité).

Intension

Chaque concept est caractérisé par un ensemble de relations qui lient chaque concept à d'autres concepts. Certaines relations simples ayant pour codomaine[6] un type de base (e.g. entier, chaîne de caractères), sont représentées par des propriétés (parfois appelées attributs). Les caractéristiques d'un concept sont l'ensemble des propriétés et relations (conditions nécessaires et suffisantes) qui décrivent ce concept.

L'intension (ou la compréhension) est l'ensemble des caractères ou propriétés contenus dans un concept et qui permettent de le définir [AN62]. Il s'agit de l'ensemble des caractéristiques qui limitent sans ambiguïté l'extension du concept (qui cible ses instances).

La définition d'un concept en intension peut être explicite ou non. La définition explicite de l'intension d'un concept se résume à l'énumération de l'ensemble des caractéristiques qui décrivent le concept.

La relation de subsomption doit être dissociée des autres relations parcequ'elle ne permet pas de caractériser directement un concept. Dans le cas où l'on définit explicitement les caractéristiques des concepts, elle peut permettre de limiter la redondance en retrouvant les caractéristiques par héritage. Si aucune caractéristique n'est définie, elle rend compte de l'inclusion des extensions et des intensions. En rendant explicite une inclusion des intensions des concepts qui se traduit par la propriété d'héritage, la relation de subsomption participe à l'intension d'un concept.

Dans la suite de ce manuscrit, l'intension d'un concept quelconque c_i d'une hiérarchie de subsomption est noté $\mathcal{I}_i$. Il s'agit d'un sous-ensemble de l'ensemble $\mathcal{I}$ des caractéristiques permettant de décrire les concepts de la hiérarchie de subsomption.

Extension

L'extension d'un concept correspond à l'ensemble des objets qui possèdent en commun les propriétés décrites en intension. On parle de l'ensemble des instances du concept. La définition d'un concept en extension peut être explicite ou non. Lorsque l'extension d'un concept est définie de manière explicite, il s'agit souvent d'un échantillon d'instances considéré comme représentatif.

Une instance d'un concept est un objet du « monde » réel qui possède toutes les caractéristiques précisées par l'intension de ce concept.

[6] Le codomaine (encore appelé range) d'une relation correspond au domaine de la relation inverse

Dans la suite de ce manuscrit, l'extension d'un concept quelconque c_i d'une hiérarchie de subsomption est noté $\mathcal{E}_i$. Il s'agit d'un sous-ensemble de l'ensemble $\mathcal{E}$ des instances du domaine considéré.

Relation de subsomption

La relation de subsomption notée $\sqsubseteq$ permet de structurer hiérarchiquement un ensemble de concepts : $c_i \sqsubseteq c_j$ signifie « c_i est plus spécifique que c_j » ou encore « c_i est subsumé par c_j ». Par définition, la relation de subsomption n'est pas stricte, ce qui signifie que c_i et c_j peuvent désigner le même concept.

La relation de subsomption est le plus souvent définie dans la littérature de manière intensionnelle ou extensionnelle :

Définition intensionnelle Un concept c_j subsume un concept c_i si tout objet décrit par c_i l'est aussi par c_j, autrement dit si l'ensemble des caractéristiques d'un objet dont la description est définie par c_i contient l'ensemble des propriétés spécifiées par c_j.

$$c_i \sqsubseteq c_j \iff \mathcal{I}_j \subseteq \mathcal{I}_i \tag{1.1}$$

Définition extensionnelle Un concept c_j subsume un concept c_i si l'ensemble des instances ciblées par c_j contient l'ensemble des instances ciblées par c_i.

$$c_i \sqsubseteq c_j \iff \mathcal{E}_i \subseteq \mathcal{E}_j \tag{1.2}$$

La relation de subsomption est donc une relation qui découle de l'observation de l'inclusion des extensions ou des intensions des concepts.

La subsomption est apparentée à la généralisation/spécialisation en représentation des connaissances par objet ou encore à l'hyperonymie/hyponymie dans le domaine de la linguistique.

L'organisation des concepts sous forme d'une hiérarchie de subsomption permet de définir chaque concept vis-à-vis de ses subsumants. Si on explicite la description intensionnelle d'un concept c_i à l'aide d'un ensemble de caractéristiques, il n'est pas nécessaire de définir toutes ses caractéristiques mais seulement celles qu'aucun de ses subsumants ne possède. L'ensemble des caractéristiques d'un concept est l'union des caractéristiques de ses subsumants. L'organisation hiérarchique permet une factorisation des caractéristiques communes au niveau des subsumants communs. On dit alors que chaque concept hérite des caractéristiques des ses subsumants. Ce mécanisme d'héritage permet donc de réduire grandement la nécessité de spécifier des caractéristiques redondantes. L'héritage est une propriété fondamentale de la relation de subsomption qui la distingue des autres relations. Lorsque deux concepts c_i et c_j subsument un même concept c_k, on parle d'héritage multiple.

Racine, feuilles

La *racine* d'une hiérarchie de subsomption est le concept qui subsume tous les autres. Il s'agit souvent d'un concept très générique (e.g. « thing », « entity », « object ») censé recouvrir toutes les instances potentielles du domaine modélisé

et que l'on nomme racine virtuelle. Lorsqu'elle ne recouvre qu'un sous-ensemble des instances du domaine considéré, on parle de racine informative. Les *feuilles* de la hiérarchie sont les concepts qui n'ont aucun subsumé.

1.3.2 Structures mathématiques

On note $\mathcal{C}$ l'ensemble des concepts de la hiérarchie de subsomption. La relation de subsomption $\sqsubseteq$ établit un ordre (ordre large) partiel sur $\mathcal{C}$: elle est *réflexive* ($\forall c_i \in \mathcal{C},\ c_i \sqsubseteq c_i$), *antisymétrique* ($\forall c_i, c_j \in \mathcal{C},\ c_i \sqsubseteq c_j \wedge c_i \neq c_j \Longrightarrow \neg c_j \sqsubseteq c_i$) et *transitive* ($\forall c_i, c_j, c_k \in \mathcal{C},\ c_i \sqsubseteq c_j \wedge c_j \sqsubseteq c_k \Longrightarrow c_i \sqsubseteq c_k$).

Hiérarchie

Deux concepts c_i et c_j de $\mathcal{C}$ sont comparables si $c_i \sqsubseteq c_j$ ou $c_j \sqsubseteq c_i$. Le couple $(\mathcal{C}, \sqsubseteq)$ est un ensemble ordonné. La relation d'ordre $\sqsubseteq$ est dite totale sur $\mathcal{C}$ si et seulement si deux éléments quelconques de $\mathcal{C}$ sont toujours comparables. Si la relation d'ordre n'est pas totale on dit qu'elle est partielle.

Considérant un sous-ensemble $\mathcal{C}_i$, le concept c_g est son plus grand élément si et seulement si $c_g \in \mathcal{C}_i$ et $\forall c_x \in \mathcal{C}_i,\ c_x \sqsubseteq c_g$ (si c_g existe, il est unique). Le concept c_p est un plus petit élément de $\mathcal{C}_i$ si et seulement si $c_p \in \mathcal{C}_i$ et $\forall c_x \in \mathcal{C}_i,\ c_x \sqsubseteq c_p$ (si c_p existe, il est unique).

Définition 1.2 *Un ensemble ordonné $(\mathcal{C}, \sqsubseteq)$ est une* hiérarchie de subsomption *si et seulement si il admet un plus grand élément que l'on appelle racine de la hiérarchie. On note $\mathcal{H}$ une telle hiérarchie.*

Une hiérarchie induite par une relation de subsomption entre concepts peut respecter des propriétés supplémentaires qui contraignent sa structure. Il peut alors s'agir d'un treillis ou d'un arbre.

Remarque. En logique de description, une hiérarchie de subsomption possède un plus grand élément (appelé TOP et noté $\top$) mais aussi un plus petit élément (appelé BOTTOM et noté $\bot$). Dans le domaine de la classification automatique, le terme de hiérarchie désigne une notion plus restrictive (un arbre).

Treillis de subsomption

On dit que la partie $\mathcal{C}_i \subseteq \mathcal{C}$ est majorée par $c_u \in \mathcal{C}$ si et seulement si pour tout c_x de $\mathcal{C}_i$, $c_x \sqsubseteq c_u$. On dit que la partie $\mathcal{C}_i \subseteq \mathcal{C}$ est minorée par $c_v \in \mathcal{C}$ si et seulement si pour tout c_x de $\mathcal{C}_i$, $c_v \sqsubseteq c_x$.

La borne supérieure de $\mathcal{C}_i$ (si elle existe) est le plus petit élément de l'ensemble des majorants de $\mathcal{C}_i$. La borne inférieure de $\mathcal{C}_i$ (si elle existe) est le plus grand élément de l'ensemble des minorants de $\mathcal{C}_i$.

Définition 1.3 *Un ensemble ordonné $(\mathcal{C}, \sqsubseteq)$ est un* treillis de subsomption *si et seulement si toute paire $\{c_i, c_j\}$ de concepts (distincts) admet une borne supérieure et une borne inférieure.*

Arbre de subsomption

On appelle éléments consécutifs de l'ensemble ordonné $(\mathcal{C}, \sqsubseteq)$ deux éléments distincts c_i et c_j qui vérifient $c_i \sqsubseteq c_j$ et $(c_i \sqsubseteq c_k \sqsubseteq c_j \Longrightarrow c_i = c_k$ ou $c_j = c_k)$. On dit alors que c_i est un prédécesseur de c_j et c_j un successeur de c_i.

Définition 1.4 *Un ensemble ordonné $(\mathcal{C}, \sqsubseteq)$ est un* arbre de subsomption *si et seulement si $\mathcal{C}$ admet un plus grand élément (qui par définition n'a donc aucun successeur) et que tous les concepts de $\mathcal{C}$ autres que son plus grand élément ont un et un seul successeur.*

Tout concept n'ayant aucun prédécesseur est appelé feuille tandis que les autres concepts sont des noeuds.

1.3.3 Treillis de Galois

Une hiérarchie de subsomption synthétise les connaissances d'un domaine sous une forme usuelle qui permet une interprétation souvent rapide par l'humain. Si de telles hiérarchies sont classiquement développées par un groupe d'experts du domaine, il est possible d'en obtenir une par la construction d'un treillis de Galois. Le principe de cette approche met formellement en évidence la nature de la relation de subsomption.

La notion de treillis de Galois d'une relation (ou treillis de concepts) est à la base d'une famille de méthodes de classification conceptuelle [GW99]. Developpée en sciences humaines par Barbut et Monjardet [BM70], cette approche a été étendue par Wille qui a utilisé la notion de treillis de Galois comme base de l'analyse formelle de concepts [Wil82].

Nous considérons l'ensemble $\mathcal{E}$ qui renferme toutes les instances d'un domaine et l'ensemble $\mathcal{I}$ des caractéristiques d'au moins une instance. On définit une relation binaire $\mathcal{R} \subseteq \mathcal{E} \times \mathcal{I}$ qui associe à chaque instance ses caractéristiques. Chaque élément du treillis de Galois est un concept formel $c_i = (\mathcal{E}_i, \mathcal{I}_i) \in \mathcal{P}(\mathcal{E}) \times \mathcal{P}(\mathcal{I})$ défini tel que :

$$\begin{cases} \mathcal{I}_i = f(\mathcal{E}_i) \text{ où } f(\mathcal{E}_i) = \{x \in \mathcal{I} | \forall y \in \mathcal{E}_i, y\mathcal{R}x\} \\ \mathcal{E}_i = f(\mathcal{I}_i) \text{ où } f(\mathcal{I}_i) = \{y \in \mathcal{E} | \forall x \in \mathcal{I}_i, y\mathcal{R}x\} \end{cases} \tag{1.3}$$

L'ensemble $\mathcal{C}$ de tous les concepts formels dérivés de $\mathcal{R}$ est ordonné par la relation $\sqsubseteq$ telle que :

$$c_i \sqsubseteq c_j \iff \mathcal{E}_i \subseteq \mathcal{E}_j \iff \mathcal{I}_j \subseteq \mathcal{I}_i \tag{1.4}$$

$\mathcal{T} = (\mathcal{C}, \sqsubseteq)$ est le treillis de Galois associé à la relation $\mathcal{R}$.

La mise en oeuvre de cette approche conduit à la construction de hiérarchies de subsomption généralement très denses qui nécessitent d'être simplifiées et élaguées. Les concepts obtenus de manière automatique ont l'avantage d'être formellement identifiés ; ils sont cependant souvent plus difficiles à appréhender par l'expert du domaine.

1.4 Conclusion

Notre travail est lié à l'essor des ontologies en Ingénierie des Connaissances. C'est pourquoi, nous avons entamé ce chapitre introductif par une présentation synthétique de la notion d'ontologie. Une ontologie comporte une hiérarchie de subsomption qui fait plus particulièrement l'objet de cette thèse.

Nous avons explicité la terminologie relative à la notion de hiérarchie de subsomption qui est employée dans la suite de ce manuscrit. Nous avons précisé les structures mathématiques (e.g. ensemble ordonné, treillis, arbre) qui permettent de la modéliser. Nous avons conclu ce chapitre par une présentation des treillis de Galois qui permet de mettre en évidence la relation qui lie l'extension et l'intension d'un concept.

Mesures entre concepts d'un réseau sémantique

2

Sommaire

Résumé

De plus en plus d'applications développées en Ingénierie des Connaissances requièrent l'évaluation d'une liaison entre concepts sur la base d'un réseau sémantique. De nombreuses mesures ont ainsi été définies de manière *ad hoc*. Ce chapitre constitue un état de l'art dans lequel nous reprenons quelques applications des mesures sémantiques qui soulignent leur utilisation multi-disciplinaire. Nous réunissons ensuite les principales mesures sémantiques de la littérature en les regroupant selon l'approche qu'elles emploient pour exploiter le réseau sémantique. Nous distinguons les approches qui considèrent le réseau comme un graphe et celles qui utilisent la notion de contenu informationnel.

2.1 Introduction

La problématique de la structuration des concepts au sein d'un réseau sémantique remonte aux travaux de Quillian [Qui68] et Collins [CL75] sur la mémoire sémantique humaine. Près de quatre décennies plus tard, les ontologies sont au coeur des systèmes informatiques développés en Ingénierie des Connaissances et les mesures sémantiques constituent une alternative pour exploiter ces référentiels de connaissances. De nombreuses mesures sémantiques ont été développées dans des domaines divers (e.g. la linguistique informatique, l'intelligence artificielle, la biologie) aussi bien pour des objectifs académiques qu'industriels [Res99] [BH01] [LSBG03] [CDKFZ04] [TSZ+04].

Quelques mesures [Sus93] [HSO98] [MMRV05] s'emploient à évaluer la proximité sémantique de deux concepts en utilisant plusieurs relations entre les concepts. Cependant, la majorité des propositions se restreint à la relation de subsomption. Bien que les contributions soient très diverses et souvent indépendantes les unes des autres, deux approches bien distinctes émergent de la littérature :

- la hiérarchie est considérée comme un graphe dans lequel on peut utiliser des indices combinatoires (e.g. profondeur, plus court chemin, densité) pour comparer des noeuds ;
- les concepts sont comparés sur la base du contenu informationnel, c'est à dire de la part d'information qu'ils partagent [Res93].

Dans ce chapitre, nous mettons tout d'abord en avant quelques applications nécessitant l'utilisation de mesures sémantiques, puis nous présentons en détail les principales mesures existantes. Nous insistons notamment sur les principes qui ont conduit à la définition de chacune de ces mesures et soulignons leurs caractéristiques essentielles.

2.2 Applications des mesures sémantiques

Avec le succès croissant des ontologies, de plus en plus d'applications utilisent les mesures sémantiques. Nous en présentons quelques unes pour illustrer leur utilité et la diversité des besoins auxquels elles répondent.

2.2.1 Recherche d'information

Les mesures sémantiques trouvent notamment leur application dans le domaine de la recherche d'information pour améliorer la pertinence des résultats et leur classement [LKL93] [MTR89] [KK90] [Sus93] [CDKFZ04].

Exemple 1. Sussna [Sus93] a contribué à la recherche d'information par mots-clés. Dans une base documentaire chaque document est indexé par un ensemble de mots-clés. Lors de la recherche d'un document, les mots-clés de la requête sont comparés avec les mots-clés indexant les documents pour fournir une liste des documents susceptibles d'intéresser l'utilisateur. Sussna relève deux problèmes dont souffre cette recherche classique dans une base documentaire :

- la recherche retourne des documents non pertinents parce qu'elle considère la polysémie d'un terme recherché et ne se focalise pas sur le sens attendu ;
- des documents pertinents sont oubliés parce qu'ils ne sont pas indexés par le terme recherché, mais par des termes similaires.

Sussna propose donc de passer par une phase de désambiguïsation[1] de mots lors de l'indexation des documents de manière à accroître la précision lors de la recherche documentaire. A l'issue de cette phase de désambiguïsation, chaque document est indexé par un ensemble de « synsets » (paires {mot,sens}) du réseau sémantique WordNet. Il s'agit donc de choisir la combinaison de « synsets » la plus appropriée pour indexer un document donné. Sussna définit ce qu'il appelle une distance sémantique qui exploite le réseau de relations de WordNet (e.g. synonymie, hyperonymie) pour évaluer l'éloignement de deux « synsets ». La somme des distances deux à deux est calculée pour chaque combinaison de « synsets » possible et l'on retient la combinaison pour laquelle cette somme est minimale.

Exemple 2. Destiné au web sémantique, Corese [CDKFZ04] est un moteur de recherche reposant sur une ontologie et testé dans de nombreuses applications industrielles. Il est dédié à la recherche de ressources web annotées en RDF(S)[2] grâce à un langage de requête spécifiquement adapté. Corese permet la recherche approximative en réponse à l'observation suivante : lorsqu'un utilisateur recherche par exemple une *personne* donnée travaillant sur un *sujet* donné, elle sera sans doute intéressée par une *équipe de recherche* travaillant sur le *sujet*. Corby et al. ont donc étendu le langage de requête de Corese pour permettre de fournir à l'utilisateur des résultats approximatifs. Pour cela, ils utilisent la mesure de Zhong et al.[ZZLY02] leur permettant d'exploiter la relation *rdfs:subClassOf*.

2.2.2 Traitement automatique du langage naturel

Exemple 1. Resnik qui est à l'origine de la notion de contenu informationnel d'un concept (cf. paragraphe 2.3.2) a proposé deux applications de sa mesure sémantique au problème de l'ambiguïté du langage naturel [Res99] :

1. La première concerne un cas particulier d'ambiguïté syntaxique qui implique à la fois les conjonctions de coordination et les noms composés qui sont sources de structure ambiguë en anglais. Resnik donne l'exemple de la phrase *food handling and storage procedures* qui représente soit une conjonction de *food handling* et *storage procedures*, soit fait référence à *handling and storage of food*. La similarité entre la signification des mots est un indice qui peut permettre de lever l'ambiguïté. Dans les phrases « a *television* and *radio* personality » et « a *psychologist* and sex *researcher* », il est clair que *television* et *radio* sont plus similaires que *television* et *personality* de la même manière que *psychologist* et *researcher* sont plus similaires que *psychologist* et *sex*. Resnik utilise sa mesure sémantique sur WordNet pour lever ces ambiguïtés.

[1] La désambiguïsation de mots consiste à déterminer le sens de chaque mot
[2] http ://www.w3.org/RDF/ et http ://www.w3.org/TR/rdf-schema/

2. La seconde application concerne la résolution de l'ambiguïté du sens des mots pour des groupes de mots en relation. L'objectif étant de sélectionner le sens approprié pour un nom donné sachant que celui-ci apparaît dans un contexte précis. Ce contexte contient d'autres noms dont le sens est lié avec celui du mot que l'on cherche à désambiguïser. Le principe de l'algorithme proposé par Resnik vise à trouver comme Sussna la combinaison de « synsets » la plus appropriée.

Exemple 2. Budanitsky et Hirst [BH01] ont comparé cinq mesures (Resnik [Res93], Jiang & Conrath [JC97], Lin [Lin98], Leacock & Chodorow [LC94] et celle de Hirst & St Onge [HSO98]) dans un système de détection et de correction de fautes d'orthographe. Pour cela, ils ont utilisé un correcteur de fautes d'orthographe basé sur la principe de la désambiguïsation des mots sous-jacent aux travaux de Hirst et St-Honge [HSO98]. Si un mot qui apparaît une seule fois dans un texte n'a pas de lien sémantique avec les autres mots du texte, mais qu'une variante orthographique[3] de ce mot aurait eu un lien, ils supposent une faute d'orthographe. D'après les résultats, la mesure de Jiang & Conrath est la plus performante. Par ailleurs, les auteurs confient qu'ils ne peuvent expliquer pourquoi cette mesure donne de bien meilleurs résultats que celle de Lin qui est pourtant composée des mêmes termes. L'étude théorique proposée dans cette thèse fournit des éléments de réponse à cette question.

2.2.3 Biologie

Exemple 1. Dans le domaine de la bio-informatique, Lord et al. [LSBG03] ont utilisé une mesure sémantique pour rechercher une protéine sémantiquement proche d'une protéine donnée. Dans ce domaine, les données contiennent beaucoup de connaissances, certaines sous forme de séquences qui sont annotées par la communauté. Cette annotation est rendue exploitable en machine grâce à l'utilisation d'une ontologie qui fournit l'ensemble du vocabulaire du domaine. Lord et al. utilisent Gene ontology [Con01] qui est une des plus importantes ontologies développées au sein de la communauté bioinformatique. Par le biais de ces annotations, chaque protéine est référencée par un ensemble de termes de l'ontologie. La similarité entre deux termes est évaluée avec la mesure de Resnik. La similarité entre deux protéines est évaluée en faisant la moyenne des similarités de termes deux à deux. Par ailleurs, ils ont montré que cette similarité exploitant les annotations est relativement bien corrélée avec une similarité de séquences de type distance d'édition utilisée classiquement sur les séquences.

Exemple 2. Une approche pour l'aide au consensus en anatomie pathologique utilisant une mesure sémantique est présentée dans [TSZ+04] et [SBT+06]. Le système IDEM (Images et Diagnostics par l'Exemple en Médecine) a pour objectif d'assister les experts dans la constitution de descriptions consensuelles de cas anatomopathologiques. Ce système permet la comparaison automatique d'anomalies morphologiques pour aider les utilisateurs à trouver un concensus.

[3] Les variantes orthographiques d'un mot w sont d'autres mots dérivés de w par le biais d'une insertion, d'une suppression ou de la permutation d'un caractère

Cette comparaison automatique repose sur l'exploitation d'un réseau sémantique structurant les termes de pathologie tumorale mammaire à l'aide d'une mesure sémantique. Afin de pouvoir comparer plusieurs mesures sémantiques (Leacock & Chodorow [LC94], Lin [Lin98], Jiang & Conrath [JC97]) et pouvoir valider l'organisation des termes, un plug-in pour l'éditeur d'ontologie Protégé a été développé. Ces trois mesures se sont révélées d'efficacité comparable dans cette application ; c'est donc la mesure la plus simple de Leacock & Chodorow qui a finalement été intégrée au logiciel IDEM pour assister les pathologistes.

2.2.4 Gestion des connaissances pour l'entreprise

Exemple 1. Dans un Web sémantique d'entreprise la construction de scénarii nécessite souvent des bases d'annotations (assertions à propos de ressources documentaires) distribuées. Pour gérer cette distribution Gandon et al. [GCDKG05] ont proposé une architecture et des protocoles permettant en particulier de maintenir la spécialisation des bases d'annotations quant aux sujets abordés dans leurs assertions. Pour cela, à chaque archivage d'une nouvelle annotation, une mesure basée sur la distance de Rada est utilisée.

Exemple 2. Laukkanen et Helin [LH05] utilisent la mesure de Stojanovic [SMS+01] dans un système de gestion des compétences qui permet de gérer les compétences des employés ainsi que d'autres ressources. Une ontologie OWL[4] répertorie l'ensemble des employés et leurs compétences. Le système permet de rechercher les employés ayant une compétence donnée. Dans le cas où personne ne possède la compétence requise, le système liste de manière ordonnée les employés ayant des compétences proches de celle recherchée grâce à la mesure de Stojanovic.

2.3 Les principales mesures sémantiques

L'ensemble des mesures sémantiques qui sont utilisées dans les applications que nous avons évoquées exploitent le réseau sémantique de diverses manières. Il existe cependant deux types de mesures bien distincts [Res99] :

- les mesures qui évaluent le degré de ressemblance entre deux concepts (*semantic similarity*) ;
- les mesures qui quantifient la force du lien, quelque soit sa nature, qui rapproche deux concepts (*semantic relatedness*).

La force du lien est une notion plus large que le degré de ressemblance : deux concepts qui se ressemblent sont nécessairement liés sémantiquement (e.g. voiture et vélo) tandis que deux concepts qui sont fortement liés sémantiquement ne se ressemblent par forcément (e.g. voiture et carburant). De façon duale, il s'agit parfois de l'absence de ressemblance ou de lien qui est évaluée.

Nous pouvons distinguer deux types de mesures que sont les mesures qui exploitent la structure du réseau comme un graphe et celles qui utilisent la

[4] http ://www.w3.org/2004/OWL/

notion de contenu informationnel. Nous détaillons ces différentes mesures en explicitant leur principe et en mettant l'accent sur leurs particularités.

2.3.1 Approche graphe

Certaines mesures sont basées sur la modélisation du réseau sémantique par un graphe dans lequel un noeud représente un concept c_i et un arc $c_i \xrightarrow{\mathcal{R}} c_j$ représente un couple (c_i, c_j) de la relation[5] quelconque $\mathcal{R}$. A chaque arc $c_x \xrightarrow{\mathcal{R}} c_y$ correspond un arc $c_y \xrightarrow{\mathcal{R}^{-1}} c_x$ issu de la relation réciproque.

La majorité des mesures de la littérature se restreint à l'exploitation de la hiérarchie de subsomption ce qui signifie que tous les arcs font référence à la relation de subsomption $\sqsubseteq$. Ces mesures évaluent un degré de ressemblance entre deux concepts. Les mesures qui proposent de prendre en compte d'autres relations mesurent plus largement la force du lien entre deux concepts.

Mesure de Rada et al.

La mesure sémantique proposée par Rada et al. [RMBB89] est basée sur une intuition triviale. Celle-ci postule que la longueur en nombre d'arcs $lc(c_i, c_j)$ du plus court chemin entre deux concepts d'une hiérarchie de subsomption $\mathcal{H}$ rend compte de leur absence de ressemblance d'un point de vue sémantique :

$$rada_{\mathcal{H}}(c_i, c_j) = lc(c_i, c_j) \tag{2.1}$$

Dans le cas d'une hiérarchie, il suffit que deux concepts aient un fils en commun pour avoir la même ressemblance que deux concepts frères. L'héritage multiple introduit donc un comportement assez peu intuitif de cette mesure.

Dans le cas d'un arbre de subsomption $\mathcal{A}$, l'unicité du chemin élémentaire[6] qui lie deux concepts nous permet d'en faire une représentation générale comme le montre la figure 2.1. Ce chemin passe par le concept nommé c_{ij} qui est le subsumant commun au concept c_i et c_j le plus profond[7]. Ainsi, nous pouvons reformuler cette mesure à l'aide des profondeurs p_i, p_j et p_{ij} des concepts c_i, c_j et c_{ij} :

$$rada_{\mathcal{A}}(c_i, c_j) = p_i + p_j - 2p_{ij} \tag{2.2}$$

Mesure de Leacock & Chodorow

Dans [Res95], Resnik fait référence à un papier non publié de Leacock et Chodorow [LC94] et présente leur mesure. La similarité de Leacock et Chodorow sera réellement présentée par ses auteurs dans [LC98]. Cette mesure correspond

[5] Un couple qui peut être obtenu par transitivité ne donne pas lieu à un arc
[6] Un chemin élémentaire est un chemin qui ne contient pas deux fois le même noeud
[7] la profondeur d'un concept correspond au nombre d'arcs qui le séparent de la racine

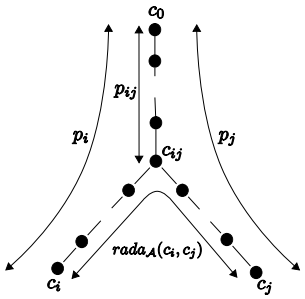


FIG. 2.1 – Mesure de Rada dans un arbre de subsomption

à une transformation de celle de Rada pour exprimer la ressemblance plutôt que l'absence de ressemblance :

$$lch_{\mathcal{H}}(c_i, c_j) = -\log\left(\frac{lc(c_i, c_j) + 1}{2p}\right) \tag{2.3}$$

où p est la profondeur[8] de la hiérarchie.

La longueur du plus court chemin entre deux concepts est normalisée grâce à une division par le double de la profondeur p de la hiérarchie. Plutôt que d'inverser l'ordre des valeurs avec une fonction linéaire, ils utilisent l'opposé du logarithme $f(x) = -\log x$. Pour éviter d'avoir le logarithme d'une valeur nulle, le plus court chemin est incrémenté de 1.

Cette mesure souffre du même problème que celle de Rada pour la prise en compte de l'héritage multiple. Si l'on se restreint à un arbre de subsomption, la reformulation est la suivante :

$$lch_{\mathcal{A}}(c_i, c_j) = -\log\left(\frac{p_i + p_j - 2p_{ij} + 1}{2p}\right) \tag{2.4}$$

Les transformations ont un impact sur la signification des mesures. La transformation effectuée ici pour satisfaire à un comportement attendu, rend l'interprétation des valeurs obtenues moins évidente.

Mesure de Zhong

Zhong et al. [ZZLY02] proposent une mesure qui évalue l'absence de ressemblance. Ils proposent également de transformer cette ressemblance à l'aide de la fonction $f(x) = 1 - x$ pour évaluer l'absence de ressemblance. Dans le cadre d'une hiérarchie restreinte à un arbre, la ressemblance qu'ils proposent est la suivante :

$$zhg_{\mathcal{A}}(c_i, c_j) = 2 \cdot \frac{1}{2^{p_{ij}+1}} - \frac{1}{2^{p_i+1}} - \frac{1}{2^{p_j+1}} \tag{2.5}$$

La fonction strictement décroissante $f(x) = 2^{-(x+1)}$ est appliquée sur les profondeurs. Si on la remplace par $f(x) = -x$, on retrouve la formule de la

[8] La profondeur de la hiérarchie correspond à la profondeur maximale d'un concept de la hiérarchie

mesure de Rada. En effet, la mesure de Zhong et al. est sensible comme celle de Rada à la taille du chemin qui sépare deux concepts. Mais lorsque la profondeur des concepts est plus importante, la distance diminue. Zhong et al. ont en effet remarqué que deux concepts séparés par un même nombre d'arcs sont plus similaires lorsqu'ils sont plus profonds (mis en évidence également par Sussna [Sus93]).

Pour adapter leur mesure à une hiérarchie, ils généralisent la profondeur p_i d'un concept c_i en prenant la longueur du plus long chemin $ll(c_i, c_0)$ de c_i jusqu'à la racine. Deux subsumants communs n'étant plus nécessairement subsumants l'un de l'autre, l'unicité de c_{ij} n'est plus garantie. Le concept c_{ij} désigne désormais l'un des subsumants communs dont le chemin à la racine est de longueur maximale. On obtient donc la formule suivante :

$$zhg_{\mathcal{H}}(c_i, c_j) = 2 \cdot \frac{1}{2^{ll(c_{ij},c_0)+1}} - \frac{1}{2^{ll(c_i,c_0)+1}} - \frac{1}{2^{ll(c_j,c_0)+1}} \tag{2.6}$$

Mesure de Wu & Palmer

La mesure proposée par Wu et Palmer [WP94] est définie seulement sur un arbre de subsomption. Elle évalue la ressemblance entre deux concepts sans recourir à une quelconque transformation du plus court chemin :

$$wup_{\mathcal{A}}(c_i, c_j) = \frac{2(p_{ij}+1)}{(p_i+1)+(p_j+1)} \tag{2.7}$$

Même si cela n'est pas indiqué par Wu et Palmer, nous remarquons que cette mesure a la forme du coefficient de Dice[9] [Dic45].

En considérant ainsi le subsumant commun le plus profond, cette mesure prend en compte le principe selon lequel deux concepts séparés par un même nombre d'arcs se ressemblent plus lorsqu'ils sont plus profonds. En effet, Si on fixe le nombre d'arcs séparant les deux concepts $(p_i+1)+(p_j+1)-2(p_{ij}+1) = constante$, la similarité croît en fonction de p_{ij} :

$$wup_{\mathcal{A}}(c_i, c_j) = \frac{2(p_{ij}+1)}{constante + 2(p_{ij}+1)}$$

La mesure de Wu & Palmer est présentée dans la littérature par Lin [Lin98] mais également par Resnik [Res99] qui introduit involontairement une variante de cette mesure lors d'un changement de notation :

$$wplr_{\mathcal{A}}(c_i, c_j) = \frac{2p_{ij}}{p_i + p_j} \tag{2.8}$$

En conséquence, deux concepts qui n'ont comme subsumant commun que la racine ont désormais une ressemblance nulle. Cette variante apparaît plus adaptée lorsque l'on a une racine virtuelle du type « entity » ou « thing ». Toutefois, ce sont les besoins de l'application qui doivent guider le choix de la mesure originale ou de sa variante.

[9] Le coefficient de Dice $\frac{2\cdot|A\cap B|}{|A|+|B|}$ définit la ressemblance de deux ensembles A et B.

Mesure de Stojanovic et al.

Stojanovic et al. [SMS+01] proposent une mesure pour évaluer la ressemblance entre deux concepts adaptée à une hiérarchie :

$$\begin{array}{l} sto_{\mathcal{H}}(c_i, c_j) = \frac{|sup(c_i) \cap sup(c_j)|}{|sup(c_i) \cup sup(c_j)|} \\ \text{avec,} \quad sup(c_x) = \{c_y | c_x \sqsubseteq c_y\} \end{array} \tag{2.9}$$

Nous remarquons que cette mesure suit la forme caractéristique du coefficient de Jaccard[10] [Jac01]. Lorsque la hiérarchie est restreinte à un arbre, le nombre de subsumants non stricts d'un concept correspond alors à sa profondeur incrémentée d'une unité :

$$sto_{\mathcal{A}}(c_i, c_j) = \frac{p_{ij} + 1}{(p_i + 1) + (p_j + 1) - (p_{ij} + 1)} \tag{2.10}$$

La proposition de Stojanovic et al. met en avant une forme généralisée de la notion de profondeur pour prendre en compte l'héritage multiple.

Mesure de Sussna

Quelques propositions ont été faites avec l'objectif de prendre en considération toutes les relations potentielles entre concepts et non plus seulement la relation de subsomption. Pour cela, Sussna [Sus93] propose de calculer une pondération $ww(c_x, c_y)$ pour chaque arc $c_x \xrightarrow{\mathcal{R}} c_y$ qui compose le plus court chemin entre deux concepts c_i et c_j avant d'en faire la somme.

La pondération $ww(c_x, c_y)$ repose sur le calcul des poids $w(c_x \xrightarrow{\mathcal{R}} c_y)$ et $w(c_y \xrightarrow{\mathcal{R}^{-1}} c_x)$. A chaque relation $\mathcal{R}$ est associé un intervalle de valeur $[min_{\mathcal{R}}; max_{\mathcal{R}}]$ qui borne les poids $w(c_x \xrightarrow{\mathcal{R}} c_y)$. Chaque poids $w(c_x \xrightarrow{\mathcal{R}} c_y)$ dépend également du nombre d'arcs $c_x \xrightarrow{\mathcal{R}} c_z$ noté $d_{\mathcal{R}}(c_x)$ et qualifié de densité locale :

$$w(c_x \xrightarrow{\mathcal{R}} c_y) = max_{\mathcal{R}} - \frac{max_{\mathcal{R}} - min_{\mathcal{R}}}{n_{\mathcal{R}}(c_x)} \tag{2.11}$$

La pondération $ww(c_x, c_y)$ est la moyenne arithmétique des poids associés aux deux arcs $c_x \xrightarrow{\mathcal{R}} c_y$ et $c_y \xrightarrow{\mathcal{R}^{-1}} c_x$ divisé par la profondeur $max(p_x, p_y)$ du plus profond des deux concepts c_x et c_y :

$$ww(c_x, c_y) = \frac{w(c_x \xrightarrow{\mathcal{R}} c_y) + w(c_y \xrightarrow{\mathcal{R}^{-1}} c_x)}{2max(p_x, p_y)} \tag{2.12}$$

Notons le caractère symétrique de cette pondération ($ww(c_x, c_y) = ww(c_y, c_x)$). Concernant la division par $2max(p_x, p_y)$, Sussna parle de mise à l'échelle relativement à la profondeur (« *depth-relative scaling* ») de manière à prendre en compte l'observation que deux concepts frères sont plus fortement reliés lorsqu'ils sont plus profond dans la hiérarchie.

[10] Le coefficient de Jaccard $\frac{|A \cap B|}{|A \cup B|}$ définit la ressemblance de deux ensembles A et B.

C'est une tentative intéressante d'évaluation de la force du lien entre deux concepts qui réduit la signification d'une relation à une pondération prise dans un intervalle de valeurs.

Mesure de Hirst & St Onge

Hirst et St Onge [HSO98] classent les diverses relations du réseau sémantique WordNet (cf. chapitre 7 section 7.3) selon leur direction : horizontales, ascendantes ou descendantes. Lorsque deux relations de directions différentes sont incidentes dans le chemin que l'on considère, on parle alors naturellement d'un changement de direction. Un chemin entre deux concepts ne sera évalué que si il ne contient pas plus de cinq arcs et si il est admissible, c'est à dire qu'il respecte certaines conditions garantissant qu'il existe bien un lien entre ces deux concepts :

- aucune autre direction ne doit précéder un lien ascendant ;
- un seul changement de direction est autorisé avec l'exception suivante : il est permis d'utiliser un lien horizontal pour faire la transition d'un lien ascendant vers un lien descendant

La force du lien entre deux concepts est défini en tenant compte de la taille du chemin admissible $la(c_i, c_j)$ contenant $nc(c_i, c_j)$ changements de direction :

$$hso(c_i, c_j) = C - la(c_i, c_j) - k \cdot nc(c_i, c_j) \tag{2.13}$$

où C et k sont deux constantes. C fixe la borne supérieure et k détermine l'impact des changements de direction. La proposition de cette mesure avec la notion de chemin admissible met en évidence la difficulté de considérer un chemin comportant des relations différentes.

Ces deux dernières mesures nous rappellent que la subsomption a un rôle radicalement différent des autres relations. La relation de subsomption limite considérablement la redondance grâce à la propriété d'héritage tandis que les autres relations permettent de caractériser les concepts.

2.3.2 Approche utilisant le contenu informationnel

Les approches utilisant le contenu informationnel exploitent uniquement la relation de subsomption généralement complétée par un corpus de textes conséquent.

Mesure de Resnik

Resnik [Res93] dit qu'intuitivement deux concepts dans une hiérarchie de subsomption devraient être considérés comme similaires lorsqu'il y a un concept spécifique qui les subsume tous les deux. Si il faut remonter très loin dans la hiérarchie pour leur trouver un subsumant commun et dans un cas extrême aller jusqu'à la racine, c'est qu'ils n'ont pas grand chose en commun. La difficulté est de savoir comment mesurer la spécificité des subsumants communs pour cibler le (ou les) plus spécifique(s) d'entre eux.

Le simple fait de compter le nombre d'arcs (en prenant la profondeur) peut être trompeur parce qu'un arc peut représenter une spécialisation plus ou moins fine suivant sa position dans la hiérarchie. De plus, prendre en compte d'autres relations peut être problématique comme le montrent Morris et Hirst [MH91]. Resnik propose alors une alternative qui est de considérer le contenu informationnel d'un concept pour évaluer sa spécificité.

Resnik associe une mesure de probabilité à l'ensemble des concepts $\mathcal{C}$ de la hiérarchie à l'aide de la fonction $P : \mathcal{C} \longrightarrow [0,1]$ telle que pour tout $c_i \in \mathcal{C}$, $P(c_i)$ est la probabilité de rencontrer une instance du concept c_i. Cela implique que P est monotone lorsque l'on se déplace dans la hiérarchie : si $c_i \sqsubseteq c_j$, alors $P(c_i) \leq P(c_j)$. La probabilité de la racine est maximale : $P(c_0) = 1$. En se basant sur la théorie de l'information, le contenu informationnel $\psi(c_i)$ d'un concept c_i peut être quantifié à l'aide de l'opposé du logarithme de sa probabilité associée : $\psi(c_i) = -\log P(c_i)$.

Resnik propose donc de mesurer la ressemblance entre deux concepts c_i et c_j par le biais du contenu informationnel maximal de leurs subsumants communs :

$$res_{\mathcal{H}}(c_i, c_j) = \max_{c_x \in sup(c_i) \cap sup(c_j)} \psi(c_x) \tag{2.14}$$

Dans un arbre de subsomption, l'ensemble des subsumants communs de c_i et c_j admet un plus petit élément c_{ij}. Il s'agit donc du subsumant commun le plus spécifique, ce qui permet de simplifier l'expression de la mesure de Resnik :

$$res_{\mathcal{A}}(c_i, c_j) = \psi(c_{ij}) \tag{2.15}$$

En pratique, pour estimer ces probabilités Resnik utilise le *Brown Corpus of American English* qui est une importante collection de textes d'un million de mots allant de l'article de presse à l'article de science fiction. Il associe une fréquence à chaque concept qu'il estime à l'aide de la fréquence des noms de ce corpus. Chaque nom qui référence un concept est comptabilisé comme occurrence de ce concept et de ses subsumants. On obtient pour chaque concept c_i sa fréquence d'apparition $freq(c_i)$ et $\mathrm{P}(c_i)$ est estimé par $\widehat{\mathrm{P}}(c_i) = \frac{freq(c_i)}{N}$, avec N le nombre total de noms qui référencent au moins un concept.

Mesure de Jiang & Conrath

Jiang et Conrath [JC97] reprennent le contenu informationnel de Resnik pour évaluer l'absence de ressemblance entre deux concept c_i et c_j dans un arbre de subsomption. Ils additionnent ce que chaque concept du plus court chemin apporte vis-à-vis de son père, ce qui revient à additionner ce que c_i et c_j apportent vis-à-vis de leur subsumant commun le plus profond c_{ij} :

$$jcn_{\mathcal{A}}(c_i, c_j) = \psi(c_i) + \psi(c_j) - 2\psi(c_{ij}) \tag{2.16}$$

Avant d'aboutir à cette mesure, ils ont adopté une approche plus générale permettant de prendre en considération d'autres aspects. A la manière de Sussna, ils calculent une pondération $ww(c_x, c_y)$ pour chaque arc (tel que $c_x \sqsubseteq c_y$) qui compose le plus court chemin entre deux concepts c_i et c_j avant

d'en faire la somme. Comme le montre l'équation 2.17, chaque pondération est le produit de quatre facteurs traduisant chacun une observation différente :

(densité) Au regard de la densité du réseau, on peut observer que la densité dans certaines parties de la hiérarchie est plus forte qu'ailleurs. Selon Richardson et Smeaton [RS95], plus la densité est importante, plus la distance entre les noeuds est faible (entre deux noeuds frères comme entre un noeud fils et son père). Ceci explique le premier facteur $\left(\beta + (1-\beta)\frac{\overline{d}}{d(c_y)}\right)$ dans lequel lorsque $\beta = 0$ l'influence de la densité locale[11] $d(c_y)$ de c_y est maximale, $\overline{d}$ désignant la densité moyenne de la hiérarchie. On considère que plus c_y a de fils vis à vis des autres concepts, plus l'absence de ressemblance entre c_y et son fils c_x est faible.

(profondeur) Il peut être argumenté que l'absence de ressemblance s'amenuise à mesure que l'on descend dans la hiérarchie du fait que la différenciation se fait sur un niveau de détail de plus en plus fin. Cette influence de la profondeur est mise en place avec le second facteur $\left(\frac{p_y+1}{p_y}\right)^{\alpha}$. Lorsque $\alpha = 1$ l'influence de la profondeur devient alors maximale.

(force de la liaison) Jiang et Conrath postulent que la force de la liaison (*link strength*) représentée par un arc entre un fils c_x et son père c_y est fonction de la probabilité conditionnelle $P(c_x/c_y)$ qu'une instance appartienne au concept fils étant donné qu'elle appartient au concept père : $P(c_x/c_y) = \frac{P(c_x)}{P(c_y)}$. Le contenu informationnel de cette probabilité conditionnelle fournit une évaluation de la force de la liaison $\psi(c_x) - \psi(c_y)$ utilisée comme troisième facteur.

(type de la relation) D'autres relations que la subsomption peuvent aussi être considérées avec un impact différent sur le calcul effectué. Le dernier facteur $T_{\mathcal{R}}(c_x, c_y)$ fixe l'influence de chaque relation $\mathcal{R}$. Cependant, Jiang et Conrath ne définissent la pondération $ww(c_x, c_y)$ que entre deux concepts c_x et c_y avec $c_x \sqsubseteq c_y$, ce qui rend ici ce facteur sans intérêt.

$$ww(c_x, c_y) = \left(\beta + (1-\beta)\frac{\overline{d}}{d(c_y)}\right)\left(\frac{p_y+1}{p_y}\right)^{\alpha}(\psi(c_x) - \psi(c_y))\,T(c_x, c_y) \tag{2.17}$$

Jiang et Conrath ont cherché à positionner pour le mieux les coefficients α et β en utilisant le jeu de test de Miller et Charles [MC91]. Le facteur de profondeur influence très peu les résultats quelque soit la valeur de α. Ils émettent l'hypothèse que ce facteur soit absorbé par le facteur qui traduit la force de la liaison. En réalité, la force de la liaison n'absorbe pas la profondeur des concepts. En revanche, c'est la somme des poids des arcs qui composent le plus court chemin qui absorbe une partie de cette information.

Nous remarquons que la densité a un effet sur la force de la liaison. En effet, lorsque le nombre de fils augmente, $P(c_x/c_y)$ diminue donc $\psi(c_x) - \psi(c_y)$ augmente. C'est d'ailleurs avec ce sens de variation que Sussna prend en compte la densité locale. Cependant, cette influence de la densité locale s'oppose au facteur de densité préconisé par Jiang et Conrath.

[11] Du fait qu'on se limite ici à la relation de subsomption, la densité locale d'un concept correspond à son nombre de fils

Jiang et Conrath s'étant inspirés des principes des mesures de Rada (plus court chemin) et de Resnik (contenu informationnel) comparent les corrélations qu'ils obtiennent sur le jeu de tests de Miller et Charles. On note une légère amélioration qu'il faut cependant relativiser étant donné que c'est ce même jeu de tests qui a servi de jeu d'apprentissage pour fixer α et β.

Mesure de Lin

Lin [Lin98] propose une similarité qui fait partie des plus étudiées sur le plan théorique. Lin tient compte de l'information partagée par les deux concepts comme Resnik, mais aussi de ce qui les distingue comme Jiang et Conrath :

$$lin_{\mathcal{A}}(c_i, c_j) = \frac{2 \cdot \psi\left(c_{ij}\right)}{\psi\left(c_i\right) + \psi\left(c_j\right)} \tag{2.18}$$

On remarque que cette mesure a tout comme la mesure de Wu & Palmer la forme du coefficient de Dice. On retrouve les deux composantes de la mesure de Jiang & Conrath [JC97] avec, à la place d'une différence, un rapport. Cependant, la différence opérée dans la mesure de Jiang & Conrath annihile l'influence du contenu informationnel de c_{ij}. Seule la différence de contenu informationnel entre chaque concept et c_{ij} fait varier la mesure de Jiang & Conrath. Si l'on conserve cette différence, peu importe le contenu informationnel de c_{ij}, la mesure de Jiang & Conrath ne varie pas contrairement à celle de Lin.

Lin fait remarquer que la mesure de Wu & Palmer est un cas particulier de sa mesure. En effet, en l'absence de corpus, si l'on considère que la probabilité qu'une instance appartienne à un concept c_x sachant qu'elle appartient à son père c_y est constante ($P(c_x/c_y) = k$ avec k une constante), alors on retrouve la formule de Wu et Palmer. Plus précisément, si l'on prend $P(c_0) = 1$ comme le préconise Resnik, on retrouve la variante de la mesure de Wu & Palmer :

$$\begin{aligned} lin_{\mathcal{A}}(c_i, c_j) &= \frac{2 \cdot \log P(c_{ij})}{\log P(c_i) + \log P(c_j)} \\ &\simeq \frac{2 \cdot \log\left(\left(\frac{1}{k}\right)^{p_{ij}}\right)}{\log\left(\left(\frac{1}{k}\right)^{p_i}\right) + \log\left(\left(\frac{1}{k}\right)^{p_j}\right)} \\ &\simeq \frac{2 \cdot p_{ij}}{p_i + p_j} \\ &\simeq wplr_{\mathcal{A}}(c_i, c_j) \end{aligned}$$

Approche de Seco

Seco et al. [SVH04] ont proposé une approche alternative qui consiste à redéfinir le contenu informationnel en considérant uniquement la hiérarchie de subsomption. Cette approche évite le recours à un corpus lors du calcul du contenu informationnel ; elle se base sur l'intuition selon laquelle l'information principale extraite par l'algorithme de calcul du contenu informationnel est en grande partie inhérente à la structure hiérarchique.

$$\psi_{sec}(c_i) = \frac{\log\left(\frac{sub(c_i)+1}{|\mathcal{C}|}\right)}{\log\left(\frac{1}{|\mathcal{C}|}\right)} = 1 - \frac{log(sub(c_i)+1)}{log(|\mathcal{C}|)} \quad \text{avec,} \quad sub(c_i) = \{c_x | c_x \sqsubseteq c_i\} \tag{2.19}$$

Le dénominateur qui est équivalent à la valeur du contenu informationnel du concept le plus informatif sert de facteur de normalisation pour que les valeurs du contenu informationnel soit dans l'intervalle $[0; 1]$. Seco et al. fixent à 0 le contenu informationnel de la racine. Ils montrent par ailleurs qu'ils obtiennent des résultats comparables voire meilleurs qu'avec l'utilisation du corpus en terme de corrélation avec le jugement humain sur le jeu de tests de Miller et Charles.

Contrairement à l'approche de Seco et al. [SVH04], nous pensons qu'il n'est pas judicieux de redéfinir le contenu informationnel. Dans l'algorithme de calcul du contenu informationnel, les occurrences de chaque concept sont comptabilisées par un balayage du corpus et l'occurrence d'un concept est prise en compte également pour tous les concepts qui le subsument. Cet algorithme de construction confère des caractéristiques à la mesure de probabilité relatives à la structure de la hiérarchie. En effet, par exemple, la probabilité $\mathrm{P}(c_i)$ d'un concept c_i décroît exponentiellement en fonction de sa profondeur, et ce plus ou moins vite en fonction du nombre de fils de chacun de ses subsumants. Il est donc envisageable de proposer de nombreuses autres approximations qui ne se basent que sur les informations présentes dans la hiérarchie de subsomption (cf. chapitres 4 et 6) ou bien qui utilisent d'autres sources d'information.

2.4 Conclusion

Après avoir présenté quelques travaux qui montrent l'étendue du champ d'application des mesures sémantiques, nous avons détaillé les principales mesures existantes. Nous avons exposé diverses mesures qui modélisent le réseau sémantique par un graphe et considèrent divers aspects de la structure qui peuvent influencer l'évaluation comme le plus court chemin, la profondeur et la densité. Nous avons également mis en avant deux mesures qui exploitent d'autres relations en complément de la relation de subsomption pour mesurer non plus la ressemblance entre concepts mais la force de leur interconnexion. La difficulté d'adopter cette dernière approche nous rappelle que la subsomption a un rôle radicalement différent des autres relations.

Dans un deuxième temps, nous avons présenté la notion de contenu informationnel d'un concept introduite par Resnik [Res93]. Cette notion de contenu informationnel est selon nous la clé pour la définition des mesures sémantiques basées sur une hiérarchie de subsomption. Contrairement à l'approche de Seco et al. [SVH04], nous pensons qu'il n'est pas nécessaire de redéfinir le contenu informationnel, mais qu'il est envisageable de proposer des approximations qui ne se basent que sur les informations de type structurelles présentes dans la hiérarchie de subsomption (cf. chapitres 4 et 6).

Deuxième partie

Un cadre général pour les mesures sémantiques

Description intensionnelle et approche ensembliste

3

Sommaire

Résumé

La définition des mesures sémantiques a tendance à être fortement déconnectée des travaux issus d'autres domaines mais qui traitent également de l'évaluation d'une liaison entre deux entités. Nous faisons la démonstration à travers ce chapitre de la possibilité de réutiliser de nombreux travaux reposant sur une modélisation ensembliste. Pour cela, nous proposons de considérer un concept par le biais de son intension. Nous dressons une typologie des mesures à travers les modèles de Tversky et des familles de similarités qui regroupent de nombreux coefficients très utilisés (e.g. Jaccard, Dice). Nous traitons également de la problématique de l'évaluation de liaisons orientées (qui donne lieu à des mesures asymétriques) en s'appuyant sur les indices de règles en ECD. Nous évoquons quelques critères de classification (l'objet, la portée et la nature) qui fournissent des éléments de comparaison entre les diverses mesures présentées.

3.1 Introduction

Évaluer une liaison entre deux objets décrits par des ensembles de caracatéristiques est une problématique centrale en analyse de données qui a de nombreuses déclinaisons selon les domaines d'application (e.g. biologie, psychologie cognitive, taxonomie numérique). De nombreuses mesures ont été définies de façon *ad hoc* pour répondre à des besoins spécifiques. Certaines mesures se trouvent utilisées de manière transversale dans des domaines différents. Dans ce chapitre, nous présentons des résultats qui peuvent s'adapter au problème de l'évaluation de liaisons entre concepts d'une hiérarchie de subsomption.

En considérant une interprétation intensionnelle de la hiérarchie, nous montrons que de nombreuses mesures peuvent se définir sur les concepts d'une hiérarchie de subsomption. Nous présentons une typologie qui permet de mettre en évidence les composantes communes de nombreuses mesures. De plus, inspirés par les travaux actuels en fouille de règles en ECD, nous utilisons la notion de règle entre concepts pour fournir un autre angle d'interprétation pour l'évaluation de liaisons orientées entre concepts.

3.2 Principe et notations

Une hiérarchie de subsomption permet de représenter des contraintes sur les intensions des concepts qu'elle structure. Elle est parfois complétée par une description intensionnelle grâce à la définition de caractéristiques (propriétés, relations) attachées aux concepts. Lorsqu'une caractérisation explicite des concepts est disponible, le cardinal de l'ensemble des caractéristiques d'un concept peut contribuer à l'évaluation de l'importance de son intension.

On remarque cependant bien souvent que chaque caractéristique n'a pas la même importance ; ce dont on peut rendre compte par le biais d'une pondération associée à chacune d'elles. L'importance de l'intension d'un concept est alors généralement évaluée par la somme des pondérations associées à ses caractéristiques.

Étant donné deux concepts c_i et c_j, les importances respectives des intensions $\mathcal{I}, \mathcal{I}_i, \mathcal{I}_j$ et $\mathcal{I}_i \cap \mathcal{I}_j$ sont notées n, n_i, n_j et n_{ij} (cf. figure 3.1). On note également $n_{\bar{i}} = n - n_i$, $n_{\bar{j}} = n - n_j$, $n_{\bar{i}j} = n_j - n_{ij}$, $n_{i\bar{j}} = n_i - n_{ij}$ et $n_{\bar{i}\bar{j}} = n - n_i - n_j + n_{ij}$ les importances respectives de $\mathcal{I} - \mathcal{I}_i$, $\mathcal{I} - \mathcal{I}_j$, $\mathcal{I}_j - \mathcal{I}_i$, $\mathcal{I}_i - \mathcal{I}_j$ et $\mathcal{I} - \mathcal{I}_i - \mathcal{I}_j$.

De nombreuses mesures issus de domaines divers visant à évaluer une liaison ou une absence de liaison entre deux ensembles peuvent être utilisées dans le cadre d'une représentation intensionnelle. Nous rappelons ici les propriétés de base associées à des mesures de ressemblance/dissemblance, puis présentons une typologie d'un ensemble de mesures de la littérature.

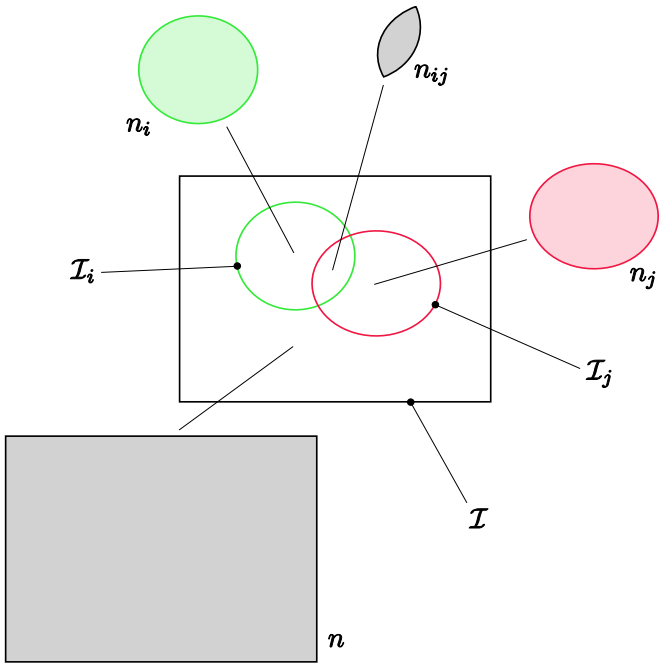


FIG. 3.1 – Diagramme de Venn des quantités observées

3.3 Les mesures de ressemblance/dissemblance

3.3.1 Quelques définitions

Définition 3.1 *Une* mesure *qui évalue une liaison entre concepts de $\mathcal{C}$ est une fonction*

$$M : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$$

Batagelj [BM95] regroupe sous le vocable anglais *resemblance* (que nous traduirons par ressemblance/dissemblance) un sous-ensemble de mesures comprenant les *backward resemblances* et les *forward resemblances* que nous traduirons respectivement par ressemblances et dissemblances (cf. définitions 3.2 et 3.3).

Définition 3.2 *Une ressemblance R est une mesure qui respecte les propriétés suivantes :*

$$R(c_i, c_j) = R(c_j, c_i) \quad \text{(symétrie)}$$
$$R(c_i, c_i) \geq R(c_i, c_j)$$

Définition 3.3 *Une dissemblance D est une mesure qui respecte les propriétés suivantes :*

$$D(c_i, c_j) = D(c_j, c_i) \quad \text{(symétrie)}$$

$$D(c_i, c_i) \leq D(c_i, c_j)$$

Souvent, les mesures de ressemblance/dissemblance respectent des propriétés supplémentaires. On peut ainsi définir les notions duales de similarité et dissimilarité.

Définition 3.4 *Une* similarité σ *est une ressemblance qui respecte les propriétés suivantes :*

$$\sigma(c_i, c_j) \geq 0 \qquad \text{(positivité)}$$
$$\sigma(c_i, c_i) = \sigma(c_j, c_j) \qquad \text{(indiscernabilité des identiques)}$$

Définition 3.5 *Une* dissimilarité δ *est une dissemblance qui respecte la propriété suivante :*

$$\delta(c_i, c_i) = 0 \qquad \text{(minimalité)}$$

Certaines dissimilarités respectent d'autres propriétés :

$$\delta(c_i, c_j) = 0 \Longrightarrow c_i = c_j \qquad \text{(identité des indiscernables)}$$
$$\delta(c_i, c_j) \leq \delta(c_i, c_k) + \delta(c_k, c_j) \qquad \text{(inégalité triangulaire)}$$

Lorsqu'elle respecte ces deux propriétés (identité des indiscernables et inégalité triangulaire), une dissimilarité est appelée distance.

3.3.2 Distance et similarité

Par définition une similarité possède une borne supérieure σ_{max} atteinte lors de la comparaison d'un concept avec lui même ($\sigma_{max} = \sigma(c_i, c_i)$). La plupart des similarités sont normalisées de sorte que $\sigma_{max} = 1$. Lorsque cela n'est pas le cas, il est possible de normaliser une similarité en posant par exemple $\sigma'(c_i, c_j) = \frac{\sigma(c_i, c_j)}{\sigma_{max}}$.

On peut transformer une similarité en dissimilarité. Les deux fonctions suivantes sont souvent utilisées :

$$\delta(c_i, c_j) = 1 - \sigma(c_i, c_j) \tag{3.1}$$

$$\delta(c_i, c_j) = \begin{cases} \frac{1}{\sigma(c_i, c_j)} & \text{si } \sigma(c_i, c_j) \neq 0 \\ +\infty & \text{sinon} \end{cases} \tag{3.2}$$

Si besoin, on vérifie le respect de l'inégalité triangulaire classique $\delta(c_i, c_j) \leq \delta(c_i, c_k) + \delta(c_k, c_j)$. Nous proposons de regarder ce que signifie le respect de cette inégalité pour la similarité lorsque celle-ci est issue de l'une des deux transformations précédentes.

Avec la transformation 3.1, cela revient à tester l'inégalité suivante :

$$\sigma(c_i, c_j) \geq \sigma(c_i, c_k) + \sigma(c_k, c_j) - 1 \tag{3.3}$$

Lorsque $\sigma(c_i, c_k) + \sigma(c_k, c_j) \leq 1$, cette inégalité n'impose aucune contrainte supplémentaire à $\sigma(c_i, c_j)$ puisque quoiqu'il arrive $\sigma(c_i, c_j) \geq 0$. Ceci est loin d'être conforme au principe de l'inégalité triangulaire classique respectée par une distance. Intuitivement, dès qu'un troisième concept c_k a une similarité non nulle à la fois avec c_i et avec c_j, l'inégalité doit imposer une similarité non nulle entre c_i et c_j.

Avec la transformation 3.2, cela revient à tester l'inégalité suivante :

$$\sigma(c_i, c_j) \geq \begin{cases} \frac{\sigma(c_i,c_k)\cdot\sigma(c_k,c_j)}{\sigma(c_i,c_k)+\sigma(c_k,c_j)} & \text{si } \sigma(c_i, c_k) + \sigma(c_k, c_j) \neq 0 \\ 0 & \text{sinon} \end{cases} \tag{3.4}$$

Lorsque $\sigma(c_i, c_k) = \sigma(c_k, c_j) = 1$, cette inégalité impose $\sigma(c_i, c_j) \geq \frac{1}{2}$. C'est une contrainte trop faible puisque dans cette situation il est clair que $\sigma(c_i, c_j) = 1$. Intuitivement, dès qu'un troisième concept c_k est identique à la fois à c_i et à c_j ($\sigma(c_i, c_k) = \sigma(c_k, c_j) = 1$), l'inégalité doit imposer une similarité maximale entre c_i et c_j.

Sur la base des deux intuitions précédentes, l'inégalité triangulaire sur une similarité doit contraindre $\sigma(c_i, c_j) \geq f(\sigma(c_i, c_k), \sigma(c_k, c_j))$ telle que :

- f est une fonction strictement croissante selon $\sigma(c_i, c_k)$ (resp. $\sigma(c_k, c_j)$) lorsque $\sigma(c_k, c_j)$ (resp. $\sigma(c_i, c_k)$) est fixe
- $f(\sigma(c_i, c_k), \sigma(c_k, c_j)) = 0$ si $\sigma(c_i, c_k) = \sigma(c_k, c_j) = 0$
- $f(\sigma(c_i, c_k), \sigma(c_k, c_j)) = 1$ si $\sigma(c_i, c_k) = \sigma(c_k, c_j) = 1$

L'adaptation de la propriété d'inégalité triangulaire pour les similarités normalisées proposée par Maguitman dans [MMRV05] a le comportement attendu :

$$\sigma(c_i, c_j) \geq \sigma(c_i, c_k) \cdot \sigma(c_k, c_j) \qquad \text{(inégalité de Maguitman)}$$

Il est alors possible de définir les deux familles de transformations réciproques qui permettent le passage d'une dissimilarité respectant l'inégalité triangulaire à une similarité respectant l'inégalité de Maguitman et vice versa :

$$\begin{aligned} &\delta(c_i, c_j) \leq \delta(c_i, c_k) + \delta(c_k, c_j) \\ &\iff -\delta(c_i, c_j) \geq -\delta(c_i, c_k) - \delta(c_k, c_j) \\ &\iff b^{-\delta(c_i,c_j)} \geq b^{-\delta(c_i,c_k)} \cdot b^{-\delta(c_k,c_j)} \end{aligned}$$

Or

$$\sigma(c_i, c_j) \geq \sigma(c_i, c_k) \cdot \sigma(c_k, c_j)$$

D'où

$$\sigma(c_i, c_j) = \begin{cases} b^{-\delta(c_i,c_j)} & \text{si } \delta(c_i, c_j) \neq +\infty \\ 0 & \text{sinon} \end{cases} \tag{3.5}$$

Et réciproquement

$$\delta(c_i, c_j) = \begin{cases} -\log_b \sigma(c_i, c_j) & \text{si } \sigma(c_i, c_j) \neq 0 \\ +\infty & \text{sinon} \end{cases} \tag{3.6}$$

Pour pouvoir effectuer une telle transformation d'une similarité en dissimilarité, on doit fixer l'échelle des valeurs de la dissimilarité en fixant la base b du logarithme. On peut définir une équivalence entre une valeur de dissimilarité δ_x et une valeur de similarité σ_x (entre leurs bornes) : $\delta_x = -\log_b \sigma_x$ d'où $b = \sqrt[\delta_x]{\sigma_x^{-1}}$.

Nous pouvons donc proposer qu'une similarité soit le pendant d'une distance lorsqu'elle respecte les propriétés supplémentaires suivantes :

$$\sigma(c_i, c_i) = 1 \qquad \text{(normalité)}$$
$$\sigma(c_i, c_j) = \sigma(c_k, c_k) \Longrightarrow c_i = c_j \qquad \text{(identité des indiscernables)}$$
$$\sigma(c_i, c_j) \geq \sigma(c_i, c_k) \cdot \sigma(c_k, c_j) \qquad \text{(inégalité de Maguitman)}$$

3.3.3 Modèles de mesures en psychologie cognitive

Les psychologues cognitifs considèrent que la similarité n'est pas nécessairement une relation symétrique et ont fourni diverses explications pour l'évaluation asymétrique de la similarité [RE04]. Cependant, nous adoptons le point de vue de Rada et al. [RMBB89] qui restreignent la similarité à l'évaluation de liaisons symétriques ; les mesures asymétriques qualifient des liaisons orientées. Nous verrons par ailleurs dans le paragraphe 3.4 l'existence de nombreux indices adaptés à l'évaluation de liaisons orientées.

En utilisant la théorie des ensembles, Tversky [Tve77] a défini deux modèles génériques de mesures en se basant sur un processus d'appariement de caractéristiques. Les mesures proposées sont fonction des caractéristiques communes aux deux concepts $\mathcal{I}_i \cap \mathcal{I}_j$ mais aussi de ce qui les différencie $\mathcal{I}_i - \mathcal{I}_j$ et $\mathcal{I}_j - \mathcal{I}_i$. Il faut noter que Tversky ne fait pas le choix d'une fonction particulière (e.g. la fonction cardinale) mais propose des modèles dont la fonction doit être instanciée. Le modèle de contraste de Tversky est défini comme suit :

$$M_{contraste}(c_i, c_j) = \theta \cdot n_{ij} - \alpha \cdot n_{i\bar{j}} - \beta \cdot n_{\bar{i}j} \tag{3.7}$$

où θ, α et $\beta \geq 0$.

Les termes α et β permettent de pondérer l'importance des caractéristiques exclusives de chaque concept relativement à leurs caractéristiques communes. Lorsque $\alpha \neq \beta$ l'ensemble des mesures est restreint aux mesures asymétriques. Restle [Res61] propose deux mesures que l'on peut rapprocher du modèle contraste. Une ressemblance $R_{restle}(c_i, c_j) = M_{contraste}(c_i, c_j) = n_{ij}$ lorsque $\theta = 1$ et $\alpha = \beta = 0$ et une dissemblance telle que $D_{restle} = -M_{contraste}(c_i, c_j) = n_{i\bar{j}} + n_{\bar{i}j}$ pour $\theta = 0$, $\alpha = \beta = 1$.

Tversky propose également le modèle ratio qui a l'avantage contrairement au précédent d'être normalisé :

$$M_{ratio}(c_i, c_j) = \frac{n_{ij}}{n_{ij} + \alpha \cdot n_{i\bar{j}} + \beta \cdot n_{\bar{i}j}} \tag{3.8}$$

où α et $\beta \geq 0$. Ce modèle généralise plusieurs mesures déjà évoquées dans la littérature en psychologie cognitive (cf. table 3.1).

La proposition de Tversky permet de définir des mesures asymétriques lorsque $\alpha \neq \beta$. Par exemple, lorsque $\alpha = 1$ et $\beta = 0$ (resp. $\alpha = 0$ et $\beta = 1$) on

Type	α	β	**Mesure**	**Références**
similarité	1	1	$\frac{n_{ij}}{n_i+n_j-n_{ij}}$	Sjöberg [Sjö72] Gregson [Gre75]
	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2\cdot n_{ij}}{n_i+n_j}$	Eisler et Ekman [EE59]
mesure	1	0	$\frac{n_{ij}}{n_i}$	Bush et Mosteller [BM51]

TAB. 3.1 – Cas particuliers du modèle ratio

obtient le degré d'inclusion de $\mathcal{I}_i$ dans $\mathcal{I}_j$ (resp. $\mathcal{I}_j$ dans $\mathcal{I}_i$). Cela correspond dans le domaine de la recherche d'information aux notions de précision et de rappel [BYN99] :

$$\mathcal{M}_{precision}(c_i, c_j) = \frac{n_{ij}}{n_i} \tag{3.9}$$

$$\mathcal{M}_{rappel}(c_i, c_j) = \frac{n_{ij}}{n_j} \tag{3.10}$$

Notons que la mesure de précision est équivalente à la mesure de Bush et Mosteller [BM51] du tableau 3.1.

La combinaison linéaire utilisée pour définir le modèle contraste et le rapport du modèle ratio rend les deux modèles génériques et complémentaires. En effet, tandis que les valeurs obtenues avec le modèle contraste dépendent de l'échelle dans laquelle sont exprimées les quantités observées (n_{ij}, $n_{i\bar{j}}$ et $n_{\bar{i}j}$), celles du modèle ratio n'en dépendent aucunement. Cette propriété constitue une différence fondamentale entre les significations de ces deux types de mesures. Les deux modèles de Tversky permettent de définir des mesures aux comportements parfois très différents et restent donc difficiles à analyser à moins de considérer des cas particuliers comme ceux que nous avons présentés.

3.3.4 Familles de similarités en analyse de données

On remarque que lorsqu'il est restreint à $\alpha = \beta \neq 0$ le modèle ratio de Tversky décrit une famille de mesures symétriques qui correspond à la famille de similarités σ_θ qui a été initialement proposée par Gower et Legendre [GL86] en analyse de données :

$$\sigma_\theta(c_i, c_j) = \frac{\theta \cdot n_{ij}}{\theta \cdot n_{ij} + n_{i\bar{j}} + n_{\bar{i}j}} \tag{3.11}$$

où $\theta \in \mathbb{R}_+^*$. Differentes valeurs de θ sont associées à des mesures usuelles (Table 3.2).

θ	**Similarité σ_θ**	**Références**
$\frac{1}{2}$	$\frac{n_{ij}}{n_{ij}+2\cdot\left(n_{i\bar{j}}+n_{\bar{i}j}\right)}$	Sokal et Sneath [SS63]
1	$\frac{n_{ij}}{n_i+n_j-n_{ij}}$	Jaccard [Jac01]
2	$\frac{2\cdot n_{ij}}{n_i+n_j}$	Dice [Dic45]

TAB. 3.2 – Association entre les valeurs de θ et σ_θ

La mesure de Eisler et Ekman (modèle ratio avec $\alpha = \beta = \frac{1}{2}$) est équivalente au coefficient de Dice tandis que celle évoquée par Sjöberg ou encore Gregson (modèle ratio avec $\alpha = \beta = 1$) correspond au coefficient de Jaccard.

Une autre famille de similarités σ_α a été proposée par Caillez et Kuntz [CK96]. Elle est définie par le rapport entre l'importance de l'intersection $\mathcal{I}_i \cap \mathcal{I}_j$ et la moyenne de Cauchy [BMV88] des importances respectives des ensembles $\mathcal{I}_i$ et $\mathcal{I}_j$:

$$\sigma_\alpha(c_i, c_j) = \frac{n_{ij}}{\mu_\alpha(n_i, n_j)} \tag{3.12}$$

où $\mu_\alpha(n_i, n_j) = \left(\frac{{n_i}^\alpha + {n_j}^\alpha}{2}\right)^{\frac{1}{\alpha}}$ pour $\alpha \in \mathbb{R}$.

La table 3.3 montre la correspondance avec des mesures connues pour diverses valeurs de α. Il est intéressant de remarquer que la mesure de Dice déjà présentée appartient également à la famille de mesures σ_α.

α	Moyenne μ_α	Similarité σ_α	Références
-1	harmonique	$\frac{1}{2}\left(\frac{n_{ij}}{n_i} + \frac{n_{ij}}{n_j}\right)$	Kulczinsky [Kul28]
0	géométrique	$\frac{n_{ij}}{\sqrt{n_i \cdot n_j}}$	Ochiaï [Och57]
1	arithmétique	$\frac{2 \cdot n_{ij}}{n_i + n_j}$	Dice [Dic45]
$-\infty$	minimum	$\frac{n_{ij}}{\min\{n_i, n_j\}}$	Simpson [Sim60]
$+\infty$	maximum	$\frac{n_{ij}}{max\{n_i, n_j\}}$	Braun-Blanquet [BB32]

Tab. 3.3 – Association entre les moyennes de Cauchy et σ_α

Gower et Legendre [GL86] ont proposé une autre famille de similarités sur le même schéma que la famille σ_θ mais qui prend également en compte les caractéristiques n'appartenant à aucun des deux concepts ($\mathcal{I} - \mathcal{I}_i - \mathcal{I}_j$) :

$$\sigma_\lambda(c_i, c_j) = \frac{\lambda \cdot \left(n_{ij} + n_{\overline{i}\overline{j}}\right)}{\lambda \cdot \left(n_{ij} + n_{\overline{i}\overline{j}}\right) + n_{i\overline{j}} + n_{\overline{i}j}} \tag{3.13}$$

où $\lambda \in \mathbb{R}_+^*$. Differentes valeurs de λ sont associées à des mesures usuelles (Table 3.4).

λ	Similarité σ_λ	Références
$\frac{1}{2}$	$\frac{n - n_{i\overline{j}} - n_{\overline{i}j}}{n + n_{i\overline{j}} + n_{\overline{i}j}}$	Rogers et Tanimoto [RT60]
1	$\frac{n_{ij} + n_{\overline{i}\overline{j}}}{n}$	Sokal et Michener [SM58]

Tab. 3.4 – Association entre les valeurs de λ et σ_λ

Il est trivial de vérifier que toutes les mesures ciblées par le modèle ratio (donc celles de la famille σ_θ) et celles de la famille σ_α ou encore σ_λ prennent leurs valeurs dans l'intervalle $[0; 1]$. Le tableau 3.5 répertorie d'autres mesures de ressemblance dont des similarités en précisant leur intervalle de définition.

La ressemblance de Russel et Rao (R_{russel}) correspond à la notion de support [AIS93] très utilisée dans les algorithmes d'extraction de règles en ECD. Les

Notation	Mesure	Intervalle	Références
R_{hamann}	$\frac{n_{ij}+n_{\bar{i}\bar{j}}-n_{\bar{i}j}-n_{i\bar{j}}}{n}$	$[-1;1]$	[Ham61]
R_{russel}	$\frac{n_{ij}}{n}$	$[0;1]$	[RR40]
R_{yule}	$\frac{n_{ij}n_{\bar{i}\bar{j}}-n_{i\bar{j}}n_{\bar{i}j}}{n_{ij}n_{\bar{i}\bar{j}}+n_{i\bar{j}}n_{\bar{i}j}}$	$[-1;1]$	[Yul00]
$R_{pearson}$	$\frac{nn_{ij}-n_in_j}{\sqrt{n_in_j\bar{n_i}\bar{n_j}}}$	$[-1;1]$	[Pea96]
$R_{michael}$	$\frac{4\left(n_{ij}n_{\bar{i}\bar{j}}-n_{i\bar{j}}n_{\bar{i}j}\right)}{\left(n_{ij}n_{\bar{i}\bar{j}}\right)^2+\left(n_{i\bar{j}}n_{\bar{i}j}\right)^2}$	$[-1;1]$	[Mic20]
R_{brin}	$\frac{nn_{ij}}{n_in_j}$	$[0;+\infty]$	[BMS97]
R_{lavrac}	$\frac{n_{ij}}{n}-\frac{n_in_j}{n^2}$	$[-0.25;0.25]$	[LFZ99]
R_{cohen}	$\frac{nn_{ij}+nn_{\bar{i}\bar{j}}-n_in_j-n_{\bar{i}}n_{\bar{j}}}{n^2-n_in_j-n_{\bar{i}}n_{\bar{j}}}$	$[-1;1]$	[Coh60]
σ_{sokal}	$\frac{1}{4}\left(\frac{n_{ij}}{n_i}+\frac{n_{ij}}{n_j}+\frac{n_{\bar{i}\bar{j}}}{n_{\bar{i}}}+\frac{n_{\bar{i}\bar{j}}}{n_{\bar{j}}}\right)$	$[0;1]$	[SS63]
σ_{sneath}	$\frac{n_{ij}n_{\bar{i}\bar{j}}}{\sqrt{n_in_j\bar{n_i}\bar{n_j}}}$	$[0;1]$	[SS63]
σ_{baroni}	$\frac{n_{ij}+\sqrt{n_{\bar{i}\bar{j}}}}{n_{ij}+n_{\bar{i}j}+n_{i\bar{j}}+\sqrt{n_{\bar{i}\bar{j}}}}$	$[0;1]$	[BUB76]

TAB. 3.5 – Autres ressemblances et similarités

mesures de ressemblance R_{brin} et R_{lavrac} sont des indices issus de L'étude des règles en ECD et traduisent respectivement les notions d'intérêt [BMS97] et de nouveauté [LFZ99] des règles. La mesure de Hamann peut s'exprimer en fonction de la mesure de Sokal et Michener [SM58] : $R_{hamann} = 2 \cdot \sigma_{\lambda=1} - 1$.

3.4 Les indices objectifs de qualité des règles

Nous considérons la notion de règle pour qualifier un type de liaison simple entre deux concepts. Cela va nous permettre d'appréhender sous un autre angle la signification des diverses mesures de ressemblance présentées dans la section 3.3. Il s'agit de décomposer chaque liaison comme une combinaison de règles entre concepts pour mieux la cerner. Cela introduit également une meilleure compréhension par comparaison des mesures entre elles.

Une règle $c_i \rightarrow c_j$ entre deux concepts c_i et c_j (cf. figure 3.2) traduit la tendance de c_j à posséder une caractéristique quand c_i la possède, et peut s'interpréter de la manière suivante : « si une caractéristique appartient à c_i alors elle appartient sûrement à c_j ». Les exemples d'une règle entre deux concepts sont les caractéristiques de $\mathcal{I}_i \cap \mathcal{I}_j$, c'est-à-dire celles qui appartiennent aux deux concepts, tandis que les contre-exemples sont les caractéristiques de $\mathcal{I}_i - \mathcal{I}_j$, celles qui appartiennent à c_i mais pas à c_j. Une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples.

À partir de deux concepts c_i et c_j il est possible de construire huit règles différentes :

$c_i \rightarrow c_j$, $\quad c_j \rightarrow c_i$,

$c_i \rightarrow \neg c_j$, $\quad c_j \rightarrow \neg c_i$,

$\neg c_i \rightarrow c_j$, $\quad \neg c_j \rightarrow c_i$,

$\neg c_i \rightarrow \neg c_j$, $\quad \neg c_j \rightarrow \neg c_i$.

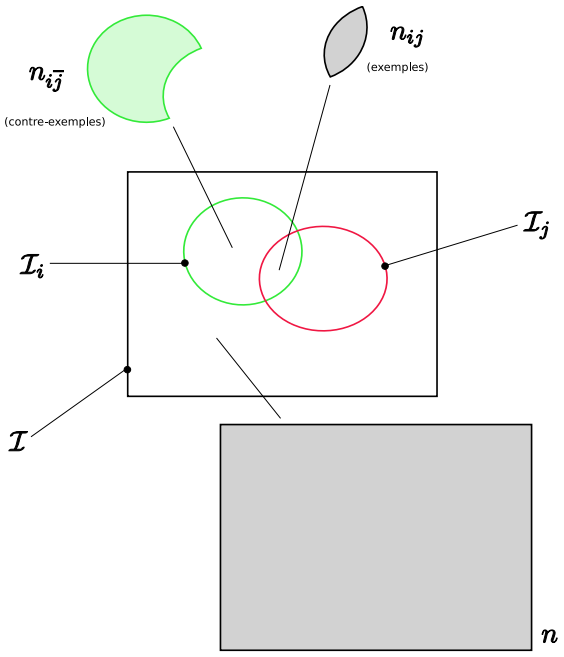


FIG. 3.2 – Diagramme de Venn pour la règle $c_i \rightarrow c_j$

Pour une règle $c_i \rightarrow c_j$, $c_i \rightarrow \neg c_j$ est la règle contraire, $c_j \rightarrow c_i$ est la règle réciproque, et $\neg c_j \rightarrow \neg c_i$ est la règle contraposée.

On retrouve une notion analogue en fouille de données. Dans ce domaine, une règle d'association entre deux sous-ensembles de caractères, de type $X \rightarrow Y$, traduit la tendance à avoir X quand on a Y. Un exemple paradigmatique est celui du panier de la ménagère. Il s'agit de découvrir des combinaisons de produits qui sont souvent achetés ensemble dans un supermarché, du type « si un client achète des huîtres, alors il achète sûrement aussi du muscadet » [Bla05].

De nombreux travaux ont été consacrés à la caractérisation de ces règles et à la construction de mesures objectives permettant d'évaluer leur pertinence. Les mesures qualifiées d' « objectives » sont calculées uniquement à partir de statistiques recueillies dans la base de données (e.g. nombre d'exemples et de contre-exemples). Elles se distinguent des mesures subjectives – que nous ne considérons pas ici – qui requièrent des informations supplémentaires (e.g. expertise des utilisateurs). Parmi les indices objectifs définis pour quantifier la qualité des règles utilisées, nous distinguons les mesures de ressemblance évoquées dans la section 3.3 et les indices asymétriques qui permettent d'évaluer des liaisons orientées. Nous montrons que ceux-ci peuvent s'interpréter dans le contexte de l'évaluation d'une liaison entre concepts dans une hiérarchie de subsomption.

Notre présentation se base sur les critères de classification proposés récemment par Blanchard et al. [Bla05]. Il s'agit d'une classification selon trois critères : l'objet de l'indice, la portée de l'indice, et la nature de l'indice. Nous

l'avons adaptée à l'évaluation de la liaison entre concepts.

3.4.1 Objet d'un indice de règle

Dans cette classification, l'objet d'un indice est la notion mesurée par celui-ci. Il peut s'agir d'un écart à l'équilibre ou d'un écart à l'indépendance.

Une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples. Ainsi, pour n_i, n_j et n donnés, la qualité de $c_i \rightarrow c_j$ est maximale lorsque $n_{ij} = \min\{n_i, n_j\}$ et minimale lorsque $n_{ij} = max\{0, n_i + n_j - n\}$. Entre ces situations extrêmes, il existe deux configurations intéressantes dans lesquelles les règles peuvent donc être considérées comme neutres ou inexistantes : l'indépendance et l'équilibre. Une règle qui se trouve dans l'une de ces configurations est à rejeter.

Indice d'écart à l'indépendance

Les concepts c_i et c_j sont indépendants si et seulement si $n_{ij} = \frac{n_i n_j}{n}$. Dans ce cas, chaque concept n'apporte aucune information sur l'autre, puisque la connaissance d'une caractéristique de l'un des concepts laisse intacte notre incertitude concernant le fait que cette caractéristique appartienne ou non à l'autre concept.

Pour deux concepts c_i et c_j donnés, il existe une unique situation d'indépendance, commune aux huit règles $c_i \rightarrow c_j$, $c_i \rightarrow \neg c_j$, $\neg c_i \rightarrow c_j$, $\neg c_i \rightarrow \neg c_j$, $c_j \rightarrow c_i$, $c_j \rightarrow \neg c_i$, $\neg c_j \rightarrow c_i$, et $\neg c_j \rightarrow \neg c_i$. Il existe deux façons de s'écarter de la situation d'indépendance :

- soit la corrélation est positive ce qui donne du poids aux quatre règles $c_i \rightarrow c_j$, $\neg c_i \rightarrow \neg c_j$, $c_j \rightarrow c_i$, et $\neg c_j \rightarrow \neg c_i$;
- soit la corrélation est négative ce qui donne du poids aux quatre règles contraires $c_i \rightarrow \neg c_j$, $\neg c_i \rightarrow c_j$, $c_j \rightarrow \neg c_i$, et $\neg c_j \rightarrow c_i$.

L'indépendance se définit à l'aide des quatre paramètres n_{ij}, n_i, n_j, et n. Ainsi, les indices d'écart à l'indépendance sont des fonctions de ces quatre paramètres (en particulier, ils décroissent tous avec n_j).

Définition 3.6 Un indice de règle I mesure un **écart à l'indépendance** si et seulement si l'indice prend une valeur fixe à l'indépendance ($n_{ij} = \frac{n_i n_j}{n}$)

Un indice d'écart à l'indépendance est utile pour découvrir des liaisons entre c_i et c_j (l'appartenance d'une caractéristique à c_i influence-t-elle son appartenance à c_j ?). Une règle $c_i \rightarrow c_j$ avec un bon écart à l'indépendance signifie : « c_j possède plus de caractéristiques de c_i qu'à l'accoutumée ».

Indice d'écart à l'équilibre

L'équilibre d'une règle $c_i \rightarrow c_j$ est défini comme la situation où la règle possède autant d'exemples que de contre-exemples : $\mathbf{n_{ij} = n_{i\bar{j}} = \frac{1}{2} n_i}$ [BGGB04a] [BGBG05]. Dans cette situation, il y a autant de caractéristiques de c_i qui sont

des caractéristiques de c_j que celles qui n'en sont pas. Une règle $c_i \rightarrow c_j$ à l'équilibre est donc autant orientée vers c_j que vers $\neg c_j$.

L'équilibre ne se définit qu'à l'aide des paramètres n_{ij} et n_i. Ainsi, les indices d'écart à l'équilibre sont généralement des fonctions de ces deux paramètres uniquement. Une règle $c_i \rightarrow c_j$ avec un bon écart à l'équilibre signifie : « c_j possède beaucoup de caractéristiques de c_i ».

Définition 3.7 Un indice de règle I mesure un **écart à l'équilibre** si et seulement si l'indice prend une valeur fixe à l'équilibre ($n_{ij} = \frac{n_i}{2}$)

Un indice d'écart à l'équilibre est utile pour prendre des décisions sur l'appartenance ou non d'une caractéristique à c_j (considérant les caractéristiques de c_i, celles-ci appartiennent elles à c_j ou pas ?). Une règle $c_i \rightarrow c_j$ avec un bon écart à l'équilibre signifie : « c_j possède la plupart des caractéristiques de c_i ».

3.4.2 Portée d'un indice de règle

La portée d'un indice est l'entité concernée par le résultat de la mesure, c'est-à-dire le type de liaison qui est évalué ; il peut s'agir d'évaluer la pertinence d'une règle, d'une règle et de sa contraposée (quasi-implication), d'une règle et de sa réciproque (quasi-conjonction), d'une règle, de sa contraposée, de sa réciproque et de la réciproque de sa contraposée (quasi-équivalence).

Règle

Une règle traduit uniquement la tendance de la conclusion à être vraie quand la prémisse est vraie. La règle $c_i \rightarrow c_j$ signifie donc « une caractéristique de c_i est une caractéristique de c_j ». Un indice de règle au sens strict I, associe au couple (c_i, c_j) la valeur $I(c_i, c_j)$ qui rend compte de la qualité de la règle $c_i \rightarrow c_j$. En d'autres termes, un indice de règle au sens strict I, associe au couple (c_i, c_j) la valeur $I(c_i, c_j)$ qui peut être interprétée comme le degré de vérité de la proposition « une caractéristique de c_i est une caractéristique de c_j ».

La règle $c_i \rightarrow c_j$ ainsi évaluée est un type de liaison simple avec une interprétation aisée. De nombreux indices de règle sont des indices de règle composés qui évaluent des liaisons plus élaborées (quasi-implication, quasi-conjonction ou quasi-équivalence).

Quasi-implication

Dans le cas général, une règle $c_i \rightarrow c_j$ n'est pas équivalente à sa contraposée $\neg c_j \rightarrow \neg c_i$. En revanche, tout comme l'implication logique, une quasi-implication notée $c_i \Rightarrow c_j$ est équivalente à sa contraposée $\neg c_j \Rightarrow \neg c_i$. La quasi-implication $c_i \Rightarrow c_j$ signifie « une caractéristique de c_i est une caractéristique de c_j et une caractéristique que n'a pas c_j n'est pas une caractéristique de c_i ».

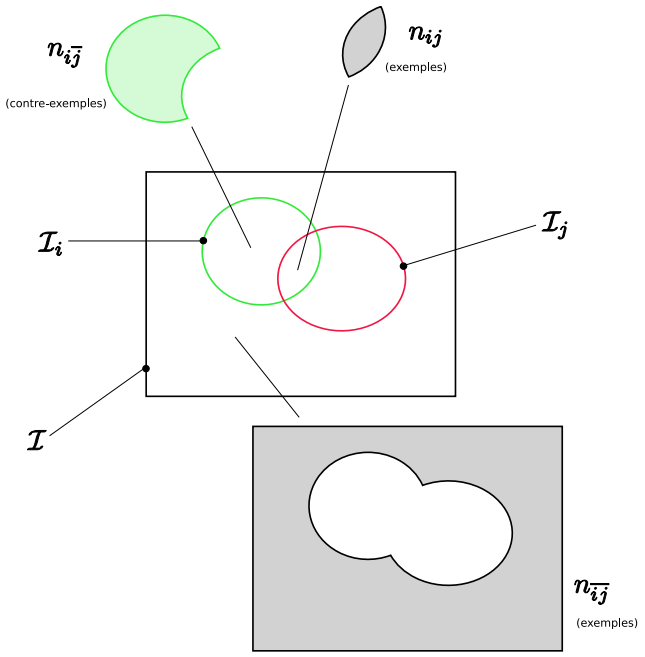


FIG. 3.3 – Diagramme de Venn pour la quasi-implication $c_i \Rightarrow c_j$

Définition 3.8 Une **quasi-implication** entre concepts est un couple (c_i, c_j) noté $c_i \Rightarrow c_j$. Les exemples d'une quasi-implication sont les caractéristiques de $\mathcal{I}_i \cap \mathcal{I}_j$ et de $\mathcal{I} - \mathcal{I}_i - \mathcal{I}_j$, tandis que les contre-exemples sont les caractéristiques de $\mathcal{I}_i - \mathcal{I}_j$. Une quasi-implication est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples. (cf. figure 3.3)

Les valeurs $I(c_i, c_j)$ prises par un indice de quasi-implication doivent donc rendre compte de la qualité des deux règles $c_i \rightarrow c_j$ et $\neg c_j \rightarrow \neg c_i$ à la fois. Un indice de quasi-implication I, associe au couple (c_i, c_j) la valeur $I(c_i, c_j)$ qui rend compte de la qualité de la quasi-implication $c_i \Rightarrow c_j$. En d'autres termes, un indice de quasi-implication I, associe au couple (c_i, c_j) la valeur $I(c_i, c_j)$ qui peut être interprétée comme le degré de vérité de la proposition « une caractéristique de c_i est une caractéristique de c_j et une caractéristique que n'a pas c_j n'est pas une caractéristique de c_i ».

Quasi-conjonction

Dans le cas général, une règle $c_i \rightarrow c_j$ n'est pas équivalente à sa réciproque $c_j \rightarrow c_i$. En revanche, tout comme la conjonction logique, une quasi-conjonction notée $c_i \leftrightarrow c_j$ est équivalente à sa réciproque $c_j \leftrightarrow c_i$. La quasi-conjonction $c_i \leftrightarrow c_j$ signifie « une caractéristique de c_i est une caractéristique de c_j et une caractéristique de c_j est une caractéristique de c_i ».

Définition 3.9 Une **quasi-conjonction** entre concepts est un couple (c_i, c_j) noté $c_i \leftrightarrow c_j$. Les exemples d'une quasi-conjonction sont les caractéristiques de

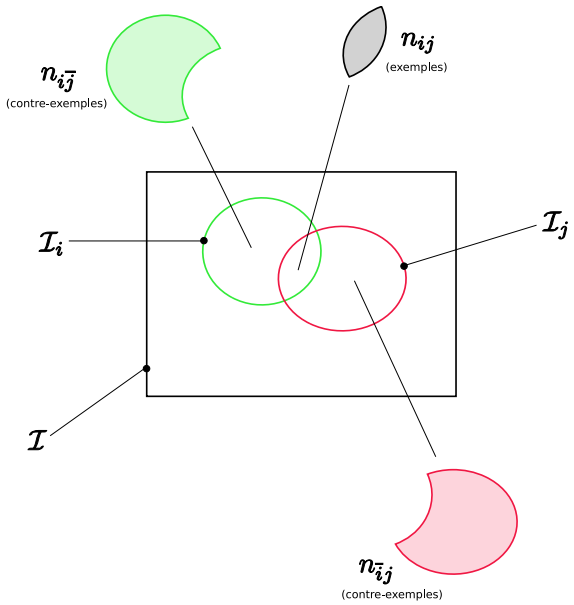


FIG. 3.4 – Diagramme de Venn pour la quasi-conjonction $c_i \leftrightarrow c_j$

$\mathcal{I}_i \cap \mathcal{I}_j$, tandis que les contre-exemples sont les caractéristiques de $\mathcal{I}_i - \mathcal{I}_j$ et $\mathcal{I}_j - \mathcal{I}_i$. Une quasi-conjonction est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples. (cf. figure 3.4)

Les valeurs $I(c_i, c_j)$ prises par un indice de quasi-conjonction doivent donc rendre compte de la qualité des deux règles $c_i \rightarrow c_j$ et $c_j \rightarrow c_i$ à la fois. Un indice de quasi-conjonction I, associe au couple (c_i, c_j) la valeur $I(c_i, c_j)$ qui rend compte de la qualité de la quasi-conjonction $c_i \leftrightarrow c_j$. En d'autre termes, un indice de quasi-conjonction I, associe au couple (c_i, c_j) la valeur $I(c_i, c_j)$ qui peut être interprétée comme le degré de vérité de la proposition « une caractéristique de c_i est une caractéristique de c_j et une caractéristique de c_j est une caractéristique de c_i ».

Quasi-équivalence

Dans le cas général, une règle $c_i \rightarrow c_j$ n'est ni équivalente à sa réciproque $c_j \rightarrow c_i$ ni équivalente à sa contraposée $\neg c_j \rightarrow \neg c_i$. En revanche, tout comme l'équivalence logique, une quasi-équivalence notée $c_i \Leftrightarrow c_j$ est équivalente à la fois à sa réciproque $c_j \Leftrightarrow c_i$ et à sa contraposée $\neg c_j \Leftrightarrow \neg c_i$ et donc à la réciproque de sa contraposée (ou contraposée de sa réciproque) $\neg c_i \Leftrightarrow \neg c_j$. La quasi-équivalence $c_i \Leftrightarrow c_j$ signifie « une caractéristique de c_i est une caractéristique de c_j et une caractéristique de c_j est une caractéristique de c_i et une caractéristique que n'a pas c_j n'est pas une caractéristique de c_i et une caractéristique que n'a pas c_i n'est pas une caractéristique de c_j ».

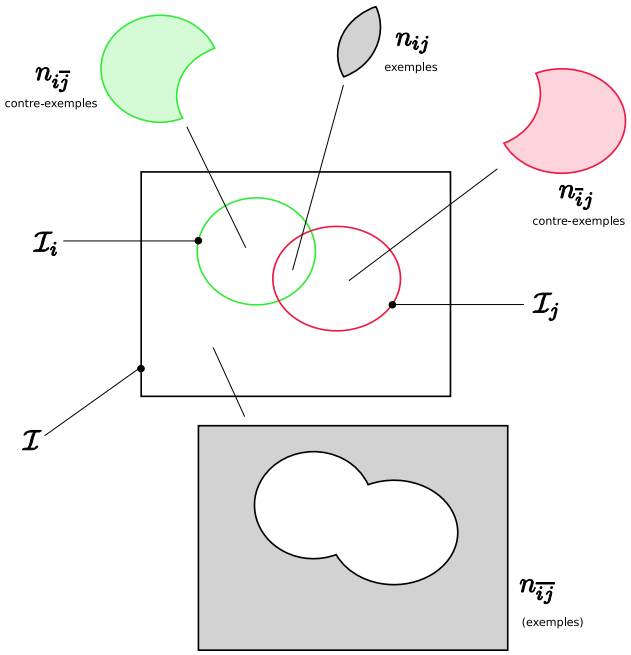


FIG. 3.5 – Diagramme de Venn pour la quasi-équivalence $c_i \Leftrightarrow c_j$

Définition 3.10 Une **quasi-équivalence** entre concepts est un couple (c_i, c_j) noté $c_i \Leftrightarrow c_j$. Les exemples d'une quasi-équivalence sont les caractéristiques de $\mathcal{I}_i \cap \mathcal{I}_j$ et de $\mathcal{I} - \mathcal{I}_i - \mathcal{I}_j$, tandis que les contre-exemples sont les individus de $\mathcal{I}_i - \mathcal{I}_j$ et $\mathcal{I}_j - \mathcal{I}_i$. Une quasi-équivalence est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples (cf. figure 3.5).

Les indices de quasi-équivalence sont à la fois des indices de quasi-implication et des indices de quasi-conjonction. Les valeurs $I(c_i, c_j)$ prises par un indice de quasi-équivalence doivent donc rendre compte de la qualité des quatre règles $c_i \rightarrow c_j$, $c_j \rightarrow c_i$, $\neg c_j \rightarrow \neg c_i$ et $\neg c_i \rightarrow \neg c_j$ à la fois. Un indice de quasi-équivalence I, associe au couple (c_i, c_j) la valeur $I(c_i, c_j)$ qui rend compte de la qualité de la quasi-équivalence $c_i \Leftrightarrow c_j$. En d'autres termes, un indice de quasi-conjonction I, associe au couple (c_i, c_j) la valeur $I(c_i, c_j)$ qui peut être interprétée comme le degré de vérité de la proposition « une caractéristique de c_i est une caractéristique de c_j et une caractéristique de c_j est une caractéristique de c_i et une caractéristique que n'a pas c_j n'est pas une caractéristique de c_i et une caractéristique que n'a pas c_i n'est pas une caractéristique de c_j ».

3.4.3 Nature d'un indice de règle

Le dernier critère de classification est la nature descriptive ou statistique des indices de règle. Les **indices descriptifs** ne varient pas lorsque n_{ij}, n_i, n_j et n sont multipliés par une constante strictement positive contrairement aux **indices statistiques**. L'utilisation d'indices statistiques en ECD permet

de prendre en compte l'augmentation de la fiabilité des résultats avec l'augmentation du volume de données. Lorsqu'il s'agit d'évaluer une règle entre concepts sur la base de leur description intensionnelle, la fonction (e.g. cardinal, somme pondérée) qui donne l'importance des caractéristiques fournit des résultats exploitables en terme de proportion uniquement et le caractère statistique de certains indices utilisés en ECD n'est pas transférable ici.

3.4.4 Classification

De par son caractère symétrique, une ressemblance est soit un indice de quasi-conjonction soit un indice de quasi-équivalence comme le montre le tableau 3.6. Les mesures de ressemblance qui sont en **gras** dans le tableau 3.6 sont des indices d'écarts à l'indépendance. Nous pouvons remarquer qu'une seule des mesures de ressemblance au sens strict répertoriées dans ce tableau n'est pas un indice d'écart à l'indépendance (R_{hamann}) et à l'inverse, qu'une seule des mesures de similarité est un indice d'écart à l'indépendance (σ_{sokal}).

Quasi-conjonction	**Quasi-équivalence**
σ_θ	R_{hamann}
σ_α	σ_λ
σ_{baroni}	σ_{sneath}
$\boldsymbol{R_{russel}}$	$\boldsymbol{R_{pearson}}$
$\boldsymbol{R_{brin}}$	$\boldsymbol{\sigma_{sokal}}$
	$\boldsymbol{R_{lavrac}}$
	$\boldsymbol{R_{cohen}}$
	$\boldsymbol{R_{michael}}$
	$\boldsymbol{R_{yule}}$

TAB. 3.6 – Classification des mesures de ressemblance

Si dans la suite nous ne traitons que des mesures de ressemblance, il est toutefois intéressant de mettre en avant l'existence des indices orientés classés dans [Bla05] comme le montre le tableau 3.7. L'analogie proposée au chapitre 5 permet d'adapter ces mesures avec une signification qui reste à étudier.

Portée / Objet	**Règle**	**Quasi-implication**
Ecart à l'équilibre	confiance [AIS93] indice de Sebag et Schoenauer [SS88] taux des exemples et contre-exemples [Gui04] estimateur laplacien de probabilité conditionnelle [BA99] indice de Ganascia [Gan91] moindre-contradiction [Aze03] TI [BGGB04b]	indice d'inclusion [GCB$^+$04] TIC [BGGB04b]
Ecart à l'indépendance	multiplicateur de cotes [LT04]	indice de Loevinger [Loe47] conviction [BMUT97]

TAB. 3.7 – Classification des indices asymétriques

3.5 Conclusion

Dans ce chapitre, nous avons analysé la pertinence de la transposition de mesures définies dans différents domaines (psychologie cognitive, taxonomie numérique, fouille de données) à l'analyse de la proximité/éloignement de concepts décrits de façon intensionnelle. Cette étude nous a permis de mettre en évidence des relations entre des mesures définies dans des contextes applicatifs différents. De plus, les différents critères de classification évoqués (objet, portée et nature) sont autant d'éléments qui ciblent la signification de chaque mesure. Cette typologie des mesures nous paraît importante pour assister l'utilisateur dans le choix d'une mesure.

Pour pouvoir exploiter ces résultats en Ingénierie des Connaissances, nous devons disposer d'une caractérisation explicite (description intensionnelle) des concepts de $\mathcal{H}$ (par le biais de propriétés et relations) ainsi que d'une fonction d'évaluation de l'importance de ces ensembles de caractéristiques. Comme le précise Bisson [Bis00], la manière dont les attributs sont pondérés est tout à fait cruciale pour que la mesure soit pertinente. Cependant, cette pondération n'est pas toujours facile à définir, notamment si l'on n'a pas d'expert du domaine capable de la fournir ; aussi, l'automatisation de ce problème a fait l'objet de nombreuses recherches [Bis00]. Par ailleurs, on ne dispose pas systématiquement de cette description intensionnelle ; elle est souvent au mieux incomplète. Les chapitres suivants montrent comment l'on peut adapter ces travaux lorsque la description intensionnelle n'est pas disponible et que l'on ne dispose que de la hiérarchie de subsomption.

Contenu informationnel dans un arbre

4

Sommaire

Résumé

Le contenu informationnel d'un concept repose sur une mesure de probabilité. Resnik propose une approximation de cette probabilité qui nécessite l'utilisation d'un corpus de textes en complément de la hiérarchie de subsomption. Ce chapitre se propose de montrer comment se passer du corpus lorsque la hiérarchie est restreinte à un arbre de subsomption. Dans un premier temps, nous exposons le principe de l'interprétation extensionnelle d'un arbre de subsomption. Nous faisons le lien entre le contenu informationnel et la notion d'interprétation extensionnelle. Nous évoquons des approximations exploitant une source d'information externe (comme un corpus) avant de proposer plusieurs approximations visant à exploiter différents aspects d'un arbre de subsomption. Ces dernières ap-

proximations relèvent de deux approches duales : (1) l'approche descendante où l'on considère l'arbre depuis la racine jusqu'aux feuilles (2) l'approche ascendante où l'on considère l'arbre depuis les feuilles jusqu'à la racine. Nous proposons également des approximations qui relèvent d'une approche mixte par agrégation d'approximations ascendantes et descendantes.

4.1 Introduction

Devant les difficultés rencontrées pour disposer d'une description intensionnelle exploitable, on se restreint souvent à l'exploitation de la hiérarchie de subsomption. Pour cela, on peut faire une interprétation extensionnelle de cette hiérarchie qui consiste à considérer les contraintes d'inclusion entre les extensions des concepts imposées par la relation de subsomption ($c_i \sqsubseteq c_j \Longrightarrow \mathcal{E}_i \subseteq \mathcal{E}_j$). La mesure de probabilité P (avec $\mathrm{P}(c_i)$ probabilité pour une instance quelconque d'appartenir à l'extension du concept c_i) que propose Resnik [Res95] permet de modéliser de telles interprétations. Cette mesure de probabilité permet d'adapter la notion d'information de la théorie de l'information de Shanon [SW49]. Resnik parle de contenu informationnel (*information content*) d'un concept (cf. chapitre 2).

En pratique, Resnik approxime la mesure de probabilité requise en extrayant les fréquences d'occurrence des concepts dans un corpus de textes conséquent. Le corpus constitue une source d'information complémentaire qui permet à Resnik de préciser l'interprétation. Nous envisageons d'autres approximations ne nécessitant aucun corpus puisqu'elle ne considèrent que les contraintes d'inclusion imposées par la hiérarchie de subsomption sur l'extension de chacun des concepts qu'elle structure. Différentes interprétations extensionnelles exploitant divers aspects de la structure hiérarchique sont alors possibles. Nous envisageons une approche descendante dans laquelle on considère l'évolution de la distribution des instances depuis la racine jusqu'aux feuilles et de façon duale l'approche ascendante. Une approche mixte (par agrégation d'approximations) permet de tirer parti des informations exploitées par ces deux approches.

Dans ce chapitre, nous discutons de l'interprétation extensionnelle d'un arbre de subsomption et définissons les notations associées. Nous rappelons ensuite la notion clef de contenu informationnel d'un concept proposée par Resnik. Enfin, nous développons plusieurs approximations restreintes à l'exploitation d'un arbre de subsomption (une extension à une hiérarchie de subsomption est présentée au chapitre 6).

4.2 Notations

Pour les besoins de ce chapitre, nous définissons des relations supplémentaires $\sqsupseteq$, $\sqsubset$, $\sqsupset$, $\prec$, $\succ$ et $\propto$:

$$c_i \sqsupseteq c_j \iff c_j \sqsubseteq c_i$$

$\hookrightarrow$ c_j est un subsumé non strict de c_i.

$c_i \sqsupset c_j \iff [c_j \sqsubseteq c_i \wedge c_i \neq c_j]$

$\hookrightarrow$ c_j est un subsumé strict de c_i.

$c_i \sqsubset c_j \iff [c_i \sqsubseteq c_j \wedge c_i \neq c_j]$

$\hookrightarrow$ c_j est un subsumant strict de c_i.

$c_i \prec c_j \iff [c_i \sqsubset c_j \wedge \nexists c_x (c_i \sqsubset c_x \wedge c_x \sqsubset c_j)]$

$\hookrightarrow$ c_j est un père (subsumant direct) de c_i.

$c_i \succ c_j \iff [c_i \sqsupset c_j \wedge \nexists c_x (c_i \sqsupset c_x \wedge c_x \sqsupset c_j)]$

$\hookrightarrow$ c_j est un fils (subsumé direct) de c_i.

$c_i \propto c_j \iff [c_j \sqsubseteq c_i \wedge \nexists c_x (c_x \sqsubset c_j)]$

$\hookrightarrow$ c_j est une feuille (concept non spécialisé) subsumée non strictement par c_i.

De manière à pouvoir « naviguer » aisément dans l'arbre $\mathcal{A}$, nous utilisons les opérateurs du type $.^{\mathcal{R}}$ qui appliqués sur un concept c_i désignent l'ensemble des images de c_i dans la relation $\mathcal{R}$: $c_i^{\mathcal{R}} = \{c_x | c_i \ \mathcal{R} \ c_x\}$. Par exemple, $c_i^{\prec}$ désigne l'ensemble des parents (subsumants directs) de c_i.

Étant donné que nous nous limitons à un arbre de subsomption, hormis la racine qui n'a pas de parents, chaque concept c_i a un et un seul père ($|c_i^{\prec}| = 1$). Cet unique père est désigné à l'aide de la notation $c_i^{\star}$.

4.3 Notion de contenu informationnel

Nous proposons de préciser la notion de contenu informationnel introduite par Resnik [Res95] pour mieux l'exploiter en détaillant au préalable celle d'interprétation extensionnelle.

4.3.1 Interprétation extensionnelle d'un arbre de subsomption

Un arbre de subsomption permet de décrire des contraintes sur les extensions des concepts qu'il structure. Par exemple, sur la figure 4.1 l'arbre proposé impose que $\mathcal{E}_8$, $\mathcal{E}_9$ et $\mathcal{E}_{10}$ soient des sous-ensembles de $\mathcal{E}_3$. Lorsqu'il n'est pas complété par une description extensionnelle, de multiples interprétations au sens des logiques de description [Nap97] sont possibles. Nous désignons par « interprétation extensionnelle » toute description extensionnelle qui respecte les contraintes imposées par la relation de subsomption.

Nous définissons une relation d'équivalence (*réflexive*, *symétrique* et *transitive*) sur l'ensemble des interprétations extensionnelles : si les cardinaux des extensions des concepts sont proportionnels alors deux interprétations sont équivalentes. Désormais, par abus de langage, nous parlons d'interprétation extensionnelle pour désigner une interprétation quelconque d'une classe d'équivalence. Une interprétation extensionnelle d'un exemple d'arbre de subsomption (qui est repris par la suite) est partiellement explicitée par la figure 4.1.

Du fait de l'existence de nombreuses interprétations extensionnelles, toute la difficulté est de choisir une interprétation pertinente. Dans notre exemple, le choix de l'interprétation extensionnelle proposée est tout à fait arbitraire.

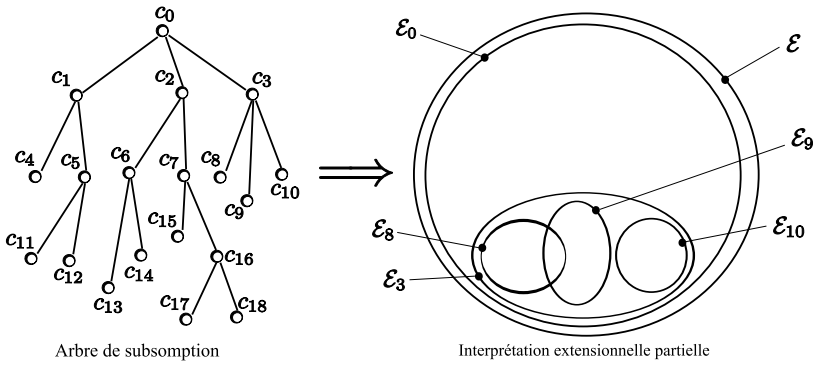


FIG. 4.1 – Interprétation extensionnelle (partielle) d'un arbre de subsomption

Pour faire une interprétation pertinente, il est possible d'utiliser des informations complémentaires externes à l'arbre de subsomption (e.g. contrainte de disjonction, échantillon d'instances, corpus de textes). En l'absence d'informations supplémentaires, une extrapolation est nécessaire de manière à aboutir à une unique interprétation extensionnelle.

L'arbre proposé dans notre exemple ne précise pas si $\mathcal{E}_8$ et $\mathcal{E}_9$ ont une intersection ou pas et le cas échéant quelle est la part de cette intersection. On peut être amené du fait d'une information externe qui le précise ou bien par hypothèse, à considérer une disjonction systématique :

$$\forall c_i, c_j \in \mathcal{C}, \neg(c_i \sqsubseteq c_j) \wedge \neg(c_j \sqsubseteq c_i) \Longrightarrow \mathcal{E}_i \cap \mathcal{E}_j = \emptyset \qquad \text{(disjonction)}$$

Lorsqu'un concept comme c_3 dans notre exemple a trois fils c_8, c_9 et c_{10}, on ne sait pas si l'ensemble des instances de c_3 se retrouve ou non dans l'extension d'au moins un de ses fils. En effet, il y a potentiellement d'autres concepts fils de c_3 qui ne sont pas définis. La complétude explicite ou supposée peut donc être envisagée de manière à préciser l'interprétation :

$$\forall c_i \in \mathcal{C}, \mathcal{E}_i = \bigcup_{c_x \in c_i^{\succ}} \mathcal{E}_x \qquad \text{(complétude)}$$

où $c_i^{\succ}$ est l'ensemble des fils de c_i.

Faire l'hypothèse de la complétude de l'arbre de subsomption, c'est considérer que l'ensemble des fils de chaque concept noeud est défini et dans ce cas, toute instance d'un concept noeud se retrouve dans un concept feuille. Si cette contrainte peut être pertinente, elle n'est pas toujours viable comme lorsque certains concepts n'ont qu'un seul fils. Il est alors envisageable de paramétrer un relâchement de cette contrainte de complétude.

La figure 4.2 illustre pour le concept c_3 et ses trois fils c_8, c_9 et c_{10} des interprétations possibles en fonction du respect ou non de la disjonction et de la complétude.

Des hypothèses sur le respect des contraintes de disjonction et de complétude ne suffisent pas à cibler une seule et unique interprétation. D'autres informations

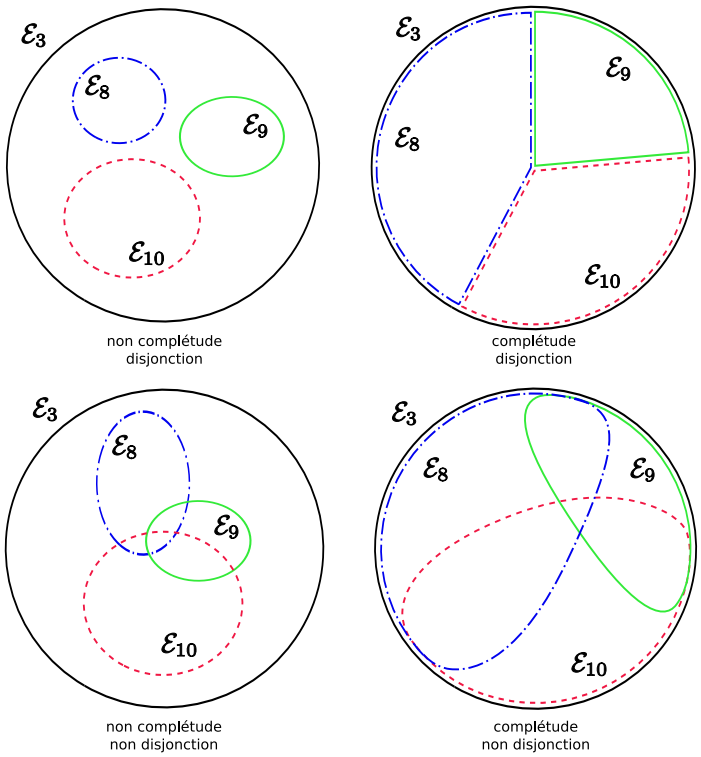


FIG. 4.2 – Diverses interprétations extensionnelles suivant le respect des contraintes de complétude et de disjonction

externes à l'arbre de subsomption ou bien des hypothèses supplémentaires sont en effet nécessaires pour cela.

Nous envisageons deux alternatives concernant l'extension de la racine de l'arbre de subsomption :

racine virtuelle L'extension de la racine couvre l'ensemble des instances du domaine

$$\mathcal{E}_0 = \mathcal{E}$$

racine informative L'extension de la racine couvre une partie seulement des instances du domaine

$$\mathcal{E}_0 \subset \mathcal{E}$$

4.3.2 Contenu informationnel d'un concept

En associant une mesure de probabilité aux concepts de l'arbre de subsomption, nous pouvons préciser une interprétation extensionnelle. L'arbre de subsomption fixe les inclusions entre les extensions des concepts tandis qu'une probabilité associée à chaque concept modélise l'importance relative de son extension.

Soit l'expérience aléatoire équiprobable $\mathcal{X}$: « On prend au hasard une instance du domaine de l'arbre de subsomption considéré ». Nous définissons $(\Omega, \mathcal{B}, \Pr)$ un espace probabilisé associé à $\mathcal{X}$ avec

$\Omega = \mathcal{E}$, l'univers des possibles,
$\mathcal{B} = \mathcal{P}(\mathcal{E})$, l'ensemble des événements (ensemble des parties de l'ensemble $\mathcal{E}$) et
Pr une mesure de probabilité sur les événements de $\mathcal{B}$

Dans cette modélisation probabiliste, $\Pr(\mathcal{E}_i)$ correspond à la probabilité pour une instance quelconque d'appartenir à l'extension $\mathcal{E}_i$ du concept c_i. Du fait de l'équiprobabilité de l'expérience aléatoire $\mathcal{X}$, $\Pr(\mathcal{E}_i) = \frac{|\mathcal{E}_i|}{|\mathcal{E}|}$. Pour une question de lisibilité, nous confondons dans la suite un concept et l'événement correspondant en parlant de la probabilité $\mathrm{P}(c_i)$ associée à un concept c_i telle que $\mathrm{P}(c_i) = \Pr(\mathcal{E}_i)$. On note également $\omega(c_i) = \left|\mathcal{E}_i - \bigcup_{c_x \in c_i^{\succ}} \mathcal{E}_x\right|$ le nombre d'instances attachées à un concept mais pas à ses fils.

Nous reprenons maintenant la notion d'information définie par Shannon dans sa théorie de l'information [SW49] sur l'espace probabilisé. Le contenu informationnel $\psi(c_i)$ d'un concept c_i correspond à la quantité d'information associée à l'événement $\mathcal{E}_i$:

$$\psi(c_i) = -\log_a \mathrm{P}(c_i) \tag{4.1}$$

où a est la base du logarithme.

L'information s'interprète comme la « quantité d'information fournie par la réalisation d'un événement ». Remarquons que celle-ci est toujours positive ou nulle et que plus un événement est improbable, plus sa réalisation apporte de l'information. De plus, le choix de la base a du logarithme permet finalement de fixer l'unité d'information utilisée. En d'autres termes, un changement de base du logarithme correspond à un changement d'échelle. La question se pose alors de savoir si il existe une base plus appropriée que les autres, c'est-à-dire qui confère au contenu informationnel d'un concept une signification aisément interprétable. Nous en discutons au cours du chapitre 5.

Si la probabilité P est associée à l'interprétation extensionnelle dans le cas fictif où l'on dispose d'une information complète sur l'ensemble des concepts, en pratique P est inconnue et le contenu informationnel repose sur une approximation $\widehat{\mathrm{P}}$.

L'approximation la plus courante est celle proposée par Resnik ; elle a été reprise ensuite dans la plupart des mesures utilisant le contenu informationnel. Cependant, nous envisageons dans la suite de ce chapitre d'autres approximations qui ne nécessitent pas de corpus additionnels –ceux-ci n'étant pas toujours disponibles– et qui exploitent la structuration combinatoire de l'arbre.

On pose $\widehat{\mathrm{P}}(c_0) = 1$ pour la racine c_0 lorsqu'elle est virtuelle. Dans le cas d'une racine informative, la probabilité $\widehat{\mathrm{P}}(c_0) < 1$ est à fixer de manière à rendre compte de la part du domaine considéré par l'arbre $\mathcal{A}$. C'est pourquoi, chaque approximation que nous définissons dans la suite de ce chapitre est exprimée en fonction de $\widehat{\mathrm{P}}(c_0)$. Nous discutons le cas échéant des valeurs pouvant être attribuées à $\widehat{\mathrm{P}}(c_0)$ et de leur signification en rapport avec l'approximation considérée.

Remarque. Dans le cas où l'arbre respecte la complétude, toutes les instances attachées à un père se retrouvent dans l'extension d'au moins l'un de ses fils. En d'autres termes, $\omega(c_i) = 0$ pour tout concept noeud c_i.

4.4 Approximations utilisant des sources d'information externes

L'approximation sur laquelle repose le contenu informationnel peut être adaptée suivant les informations disponibles. Nous analysons le cas où l'arbre de subsomption est accompagné d'un échantillon d'instances pour chacune des feuilles et le cas où il est complété par un corpus de textes.

Utilisation d'un échantillon d'instances

On dispose d'un échantillon $\widetilde{\mathcal{E}}_i$ de l'extension $\mathcal{E}_i$ de chaque concept feuille $c_i \in c_0^{\propto}$ ($c_0^{\propto}$: ensemble des concepts feuilles subsumés par c_0). Ces échantillons constituent une source d'information complémentaire qui guide les interprétations extensionnelles possibles. Pour cibler une unique interprétation, nous pouvons supposer que l'arbre est défini de manière complète (respect de la propriété de complétude). L'extension de chaque concept noeud c_i de l'arbre est alors définie en faisant l'union des extensions de ses fils ($\bigcup_{c_x \in c_i^{\succ}} \widetilde{\mathcal{E}}_x$). Par application récursive, cela correspond à l'union des extensions des feuilles subsumées par c_i ($\bigcup_{c_x \in c_i^{\propto}} \widetilde{\mathcal{E}}_x$). Pour prendre en compte le statut de la racine grâce à $\widehat{\mathrm{P}}(c_0)$, l'approximation des probabilités $\widehat{\mathrm{P}}_e$ est définie comme suit :

$$\widehat{\mathrm{P}}_e(c_i) = \widehat{\mathrm{P}}(c_0) \cdot \frac{\left| \bigcup_{c_x \in c_i^{\propto}} \widetilde{\mathcal{E}}_x \right|}{\left| \bigcup_{c_x \in c_0^{\propto}} \widetilde{\mathcal{E}}_x \right|} \tag{4.2}$$

où $c_i^{\propto}$ désigne l'ensemble des feuilles subsumées non strictement par c_i (si c_i est une feuille, $c_i^{\propto} = \{c_i\}$).

Remarque 1. Il est cependant possible de relaxer la contrainte de complétude en considérant qu'il manque un certain nombre ϵ de concept fils pour chaque chaque noeud. Une première alternative est de considérer que tous ces fils supplémentaires sont des feuilles dont la taille de l'extension est du même ordre de grandeur que celle des autres feuilles. Pour cela, nous définissons une approximation $\widehat{\omega}_e$ de ω en considérant la moyenne des nombres d'instances de chaque feuille. Nous redéfinissons ainsi l'approximation $\widehat{\mathrm{P}}_e$ en sachant qu'avec $\epsilon = 0$, on

retrouve la définition précédente :

$$\widehat{\mathrm{P}}_e(c_i) = \widehat{\mathrm{P}}(c_0) \cdot \frac{\left|\bigcup_{c_x \in c_i^{\propto}} \widetilde{\mathcal{E}}_x\right| + \sum_{c_x \in c_i^{\sqsupset} - c_i^{\propto}} \widehat{\omega}_e(c_x)}{\left|\bigcup_{c_x \in c_0^{\propto}} \widetilde{\mathcal{E}}_x\right| + \sum_{c_x \in \mathcal{C} - c_0^{\propto}} \widehat{\omega}_e(c_x)} \tag{4.3}$$
$$\text{avec } \widehat{\omega}_e(c_i) = \frac{\epsilon}{|c_0^{\propto}|} \cdot \left|\bigcup_{c_x \in c_0^{\propto}} \widetilde{\mathcal{E}}_x\right|, \text{ pour } c_i \in \mathcal{C} - c_0^{\propto}$$

où $c_i^{\sqsupset}$ est l'ensemble des concepts subsumés non strictement ($c_i \in c_i^{\sqsupset}$) par c_i.

Remarque 2. Nous remarquons cependant que l'oubli d'un fils au niveau d'un concept très général devrait avoir plus d'influence qu'au niveau d'un concept plus spécifique. En effet, on peut considérer les concepts oubliés non pas comme des feuilles mais comme des concepts racine d'un sous-arbre aussi imposant que celui de leurs frères. Nous redéfinissons l'approximation $\widehat{\mathrm{P}}_e$ en sachant qu'avec $\epsilon = 0$, on retrouve toujours un respect de la complétude :

$$\widehat{\mathrm{P}}_e(c_i) = \widehat{\mathrm{P}}(c_0) \cdot \frac{\left|\bigcup_{c_x \in c_i^{\propto}} \widetilde{\mathcal{E}}_x\right| + \sum_{c_x \in c_i^{\sqsupset} - c_i^{\propto}} \widehat{\omega}_e(c_x)}{\left|\bigcup_{c_x \in c_0^{\propto}} \widetilde{\mathcal{E}}_x\right| + \sum_{c_x \in \mathcal{C} - c_0^{\propto}} \widehat{\omega}_e(c_x)} \tag{4.4}$$
$$\text{avec } \widehat{\omega}_e(c_i) = \frac{\epsilon}{c_i^{\sqsupset}} \cdot \left(\left|\bigcup_{c_x \in c_i^{\propto}} \widetilde{\mathcal{E}}_x\right| + \sum_{c_x \in c_i^{\sqsupset} - c_i^{\propto}} \widehat{\omega}_e(c_x)\right), \text{ pour } c_i \in \mathcal{C} - c_0^{\propto}$$

où $c_i^{\sqsupset}$ est l'ensemble des concepts subsumés strictement ($c_i \notin c_i^{\sqsupset}$) par c_i .

Utilisation d'un corpus de textes

Parfois on dispose en plus de l'arbre $\mathcal{A}$ d'un corpus de textes conséquent du domaine. Resnik [Res95] propose d'extraire d'un corpus de textes l'effectif des occurrences de chaque terme qui désigne un concept de l'arbre. Pour un concept c_i, cet effectif $\widehat{\omega}_r(c_i)$ est une approximation de ω. L'approximation $\widehat{\mathrm{P}}_r$ de Resnik est la suivante :

$$\widehat{\mathrm{P}}_r(c_i) = \frac{\sum_{c_x \in c_i^{\sqsupset}} \widehat{\omega}_r(c_x)}{\sum_{c_x \in \mathcal{C}} \widehat{\omega}_r(c_x)} \tag{4.5}$$

où $c_i^{\sqsupset}$ est l'ensemble des subsumés de c_i.

En additionnant ainsi les fréquences d'occurrences de tous les subsumés, Resnik considère que les extensions de deux concepts non subsumants l'un de l'autre

et n'ayant pas de subsumés en commun sont disjointes (respect de la contrainte de disjonction). De plus, le fait de comptabiliser $\widehat{\omega}_r(c_i)$ occurrences propres à c_i (que l'on ne retrouve pas dans ses fils) constitue un relâchement de la contrainte de complétude qui est illustré sur la figure 4.3 par l'ajout d'un concept fils supplémentaire pour chaque noeud. Nous avons positionné les concepts feuilles sur la figure 4.3 de manière plus ou moins aléatoire pour représenter des fréquences d'occurrence tirées d'un corpus.

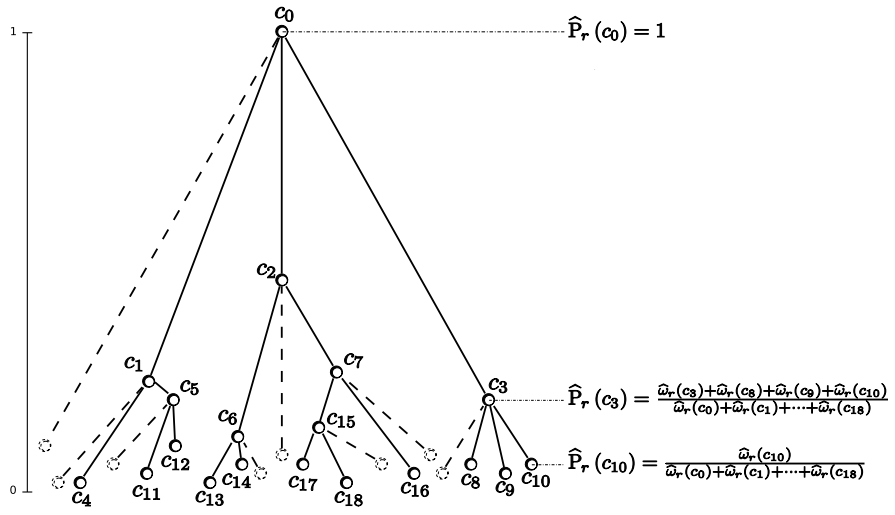


FIG. 4.3 – Application de l'approximation $\widehat{\mathrm{P}}_r$

L'approximation proposée par Resnik considère intrinsèquement la racine comme virtuelle. Pour intégrer comme précédemment la possibilité de choisir le statut de la racine nous proposons l'approximation $\widehat{\mathrm{P}}_c$:

$$\widehat{\mathrm{P}}_c(c_i) = \widehat{\mathrm{P}}(c_0) \cdot \frac{\sum_{c_x \in c_i^{\sqsupset}} \widehat{\omega}_r(c_x)}{\sum_{c_x \in \mathcal{C}} \widehat{\omega}_r(c_x)} \tag{4.6}$$

Les approximations suivantes proposent diverses alternatives pour se passer du corpus. Nous cherchons à exploiter au maximum la structure de l'arbre pour affiner l'approximation du contenu informationnel.

4.5 Approximations exploitant la structure de l'arbre de subsomption

Une tentative de redéfinition du contenu informationnel à partir de la structure de l'arbre a été proposée par Seco et al. [SVH04]. L'idée sous-jacente est que plus un concept a de subsumés, moins il apporte d'information et que les

concepts feuilles apportent par conséquent une information maximale :

$$\begin{aligned} \psi_{sec}(c_i) &= \frac{\log_a\left(\frac{\left|c_i^{\sqsupseteq}\right|}{|\mathcal{C}|}\right)}{\log_a\left(\frac{1}{|\mathcal{C}|}\right)} \\ &= 1 - \frac{\log_a\left(\left|c_i^{\sqsupseteq}\right|\right)}{\log_a(|\mathcal{C}|)} \end{aligned} \tag{4.7}$$

Pour exploiter plus largement l'information contenue dans la structure hiérarchique, nous proposons de considérer différentes hypothèses de distribution des instances sur les concepts de l'arbre. Notre réflexion a donné lieu à des approximations qui relèvent de trois approches différentes :

1. l'approche descendante
   - $\widehat{\mathrm{P}}_p$ repose sur l'hypothèse d'une réduction exponentielle des extensions avec l'augmentation de la profondeur ;
   - $\widehat{\mathrm{P}}_s$ repose sur une hypothèse d'équirépartition des instances d'un père vers ses fils, et ce depuis la racine jusqu'aux feuilles ;
2. l'approche ascendante
   - $\widehat{\mathrm{P}}_h$ repose sur l'hypothèse d'un accroissement exponentiel des extensions avec l'augmentation de la hauteur ;
   - $\widehat{\mathrm{P}}_g$ repose sur une hypothèse d'équirépartition des instances de la racine sur les feuilles et de regroupement des instances des fils pour former l'extension de leur père depuis les feuilles jusqu'à la racine.
3. l'approche mixte
   - $\widehat{\mathrm{P}}_{ph}$ agrège les approximations $\widehat{\mathrm{P}}_p$ et $\widehat{\mathrm{P}}_h$;
   - $\widehat{\mathrm{P}}_{sg}$ agrège les approximations $\widehat{\mathrm{P}}_s$ et $\widehat{\mathrm{P}}_g$.

Les approximations qui relèvent de l'approche ascendante considèrent que toutes les feuilles ont la même spécificité. De manière duale, l'approche descendante va au contraire les différencier le plus possible. C'est pourquoi nous proposons une approche mixte à travers l'agrégation d'approximations issues de ces deux approches complémentaires.

4.5.1 Approche descendante

Approximation $\widehat{\mathrm{P}}_p$

La considération d'une distribution uniforme des instances pour les concepts de même profondeur relève d'une approche descendante dans laquelle à chaque niveau de profondeur[1], le nombre d'instances des concepts par rapport au niveau précédent est divisé.

Le contenu informationnel du concept racine c_0 est minimal (nul si la racine est virtuelle) et étant donné deux concepts c_i et c_j tels que $c_i \sqsubseteq c_j$, $\psi(c_i) > \psi(c_j)$. Il y a donc un rapport entre le contenu informationnel d'un concept c_i et sa profondeur p_i :

$$p_i = \left|c_i^{\sqsubset} - \{c_0\}\right| = \left|c_i^{\sqsubset}\right| - 1 = \left|c_i^{\sqsubset}\right| \tag{4.8}$$

[1] un $n^{\text{ème}}$ niveau de profondeur contient tous les concepts de profondeur $n+1$. Le premier niveau contient donc uniquement la racine, le niveau suivant contient tous les fils de la racine et ainsi de suite.

où $c_i^{\sqsubset}$ (resp. $c_i^{\sqsubseteq}$) est l'ensemble des concepts subsumants strictement (resp. non strictement) c_i ($c_i^{\sqsubseteq} = c_i^{\sqsubset} \cup \{c_i\}$).

On fait ici l'hypothèse que le nombre de concepts par niveau de profondeur évolue exponentiellement en fonction de la profondeur : du $n^{\text{ème}}$ niveau de l'arbre au $n+1^{\text{ème}}$ niveau, le nombre de concepts est multiplié par un facteur κ ($\kappa > 1$). Le scalaire κ correspond à l'ordre de grandeur du nombre de spécialisations d'un concept sur la globalité de l'arbre.

Si l'on considère une distribution uniforme des instances pour les concepts d'un même niveau, les cardinaux des extensions des concepts du $n+1^{\text{ème}}$ niveau résultent donc de la division par κ des cardinaux des extensions des concepts du $n^{\text{ème}}$ niveau : $\forall c_i, c_j \in \mathcal{C}, p_i = p_j \Longrightarrow |\mathcal{E}_i| = |\mathcal{E}_j|$.

Dans ce cas, la probabilité qu'une instance soit associée à un concept c_i décroît exponentiellement avec la profondeur de c_i dans $\mathcal{A}$. Ainsi, on peut approximer P par $\widehat{\mathrm{P}}_p$ (cf. figure 4.4 pour l'exemple) tel que :

$$\begin{aligned} \widehat{\mathrm{P}}_p\left(c_i\right) &= \frac{\widehat{\mathrm{P}}_p\left(c_i^\star\right)}{\kappa} \\ &= \frac{\widehat{\mathrm{P}}\left(c_0\right)}{\kappa^{p_i}} \end{aligned} \tag{4.9}$$

où $c_i^\star$ désigne l'unique père de c_i et p_i la profondeur de c_i.

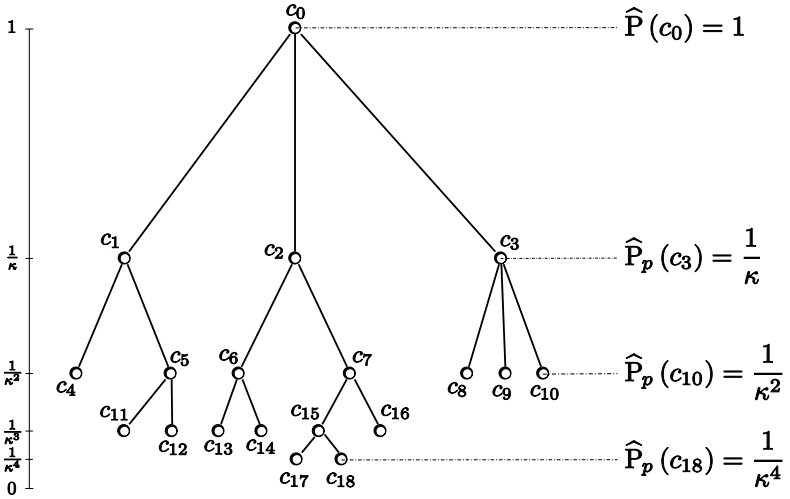


FIG. 4.4 – Application de l'approximation $\widehat{\mathrm{P}}_p$ avec $\widehat{\mathrm{P}}\left(c_0\right) = 1$

Remarquons que si l'on fixe comme unité d'information (base du logarithme) cet ordre de grandeur ($a = \kappa$), le contenu informationnel d'un concept c_i est équivalent à sa profondeur augmentée du contenu informationnel de la racine :

$$\begin{aligned} \psi_p\left(c_i\right) &= -\log_\kappa \widehat{\mathrm{P}}_p\left(c_i\right) \\ &= -\log_\kappa \frac{\widehat{\mathrm{P}}(c_0)}{\kappa^{p_i}} \\ &= p_i - \log_\kappa \widehat{\mathrm{P}}\left(c_0\right) \\ &= p_i + \psi\left(c_0\right) \end{aligned} \tag{4.10}$$

Approximation $\widehat{\mathrm{P}}_s$

Nous proposons maintenant d'étudier la répartition des instances depuis la racine jusqu'aux feuilles. Le plus simple en l'absence d'informations complémentaires est de considérer une équirépartition des instances du père sur ses fils et ce de manière disjointe :

$$\forall c_i, c_j \in \mathcal{C}, c_i^\star = c_j^\star \Longrightarrow |\mathcal{E}_i| = |\mathcal{E}_j| \qquad \text{(uniformité des frères)}$$

$$\forall c_i, c_j \in \mathcal{C}, c_i^\star = c_j^\star \Longrightarrow \mathcal{E}_i \cap \mathcal{E}_j = \emptyset \qquad \text{(disjonction entre frères)}$$

où $c_i^\star$ désigne l'unique père de c_i.

Lorsque la propriété de complétude est satisfaite, on obtient donc :

$$\widehat{\mathrm{P}}_s\left(c_i\right) = \frac{\widehat{\mathrm{P}}_s\left(c_i^\star\right)}{\left|\left(c_i^\star\right)^\succ\right|} \tag{4.11}$$

où $\left(c_i^\star\right)^\succ$ désigne l'ensemble des fils du père de c_i.

La figure 4.5 illustre le calcul de cette approximation. Par exemple, la probabilité attachée à c_3 se calcule en divisant la probabilité de son père c_0 par le nombre de fils de c_0 :

$$\widehat{\mathrm{P}}_s\left(c_3\right) = \frac{\widehat{\mathrm{P}}(c_0)}{|\{c_1, c_2, c_3\}|} = \frac{1}{3}$$

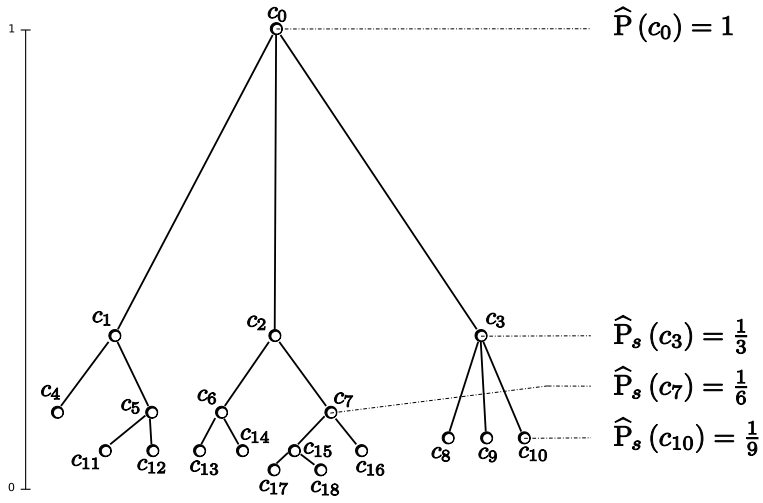


FIG. 4.5 – Application de l'approximation $\widehat{\mathrm{P}}_s$ avec $\widehat{\mathrm{P}}\left(c_0\right) = 1$

Le contenu informationnel $(-\log \widehat{\mathrm{P}}_s)$ issu de cette approximation rend compte du degré de spécificité vis-à-vis de la racine. La profondeur traduit une partie de l'information exploitée par ce degré de spécificité. Il s'agit donc d'une approximation qui affine $\widehat{\mathrm{P}}_p$ en considérant en plus pour chaque subsumant son nombre de fils.

Remarque. Il arrive parfois que la complétude ne soit pas respectée (par exemple, lorsqu'un concept n'a qu'un seul fils). Dans ce cas, on considère ϵ fils supplémentaires pour chaque concept noeud de manière à relâcher cette contrainte. Que ces concepts soient des feuilles ou bien des racines de sous-arbres aussi imposants que ceux de leurs frères, il n'y a qu'une seule possibilité de redéfinition de l'approximation $\widehat{\mathrm{P}}_s$ (cf. figure 4.6 pour l'exemple) :

$$\widehat{\mathrm{P}}_s\left(c_i\right) = \frac{\widehat{\mathrm{P}}_s\left(c_i^\star\right)}{\left|\left(c_i^\star\right)^\succ\right| + \epsilon} \tag{4.12}$$

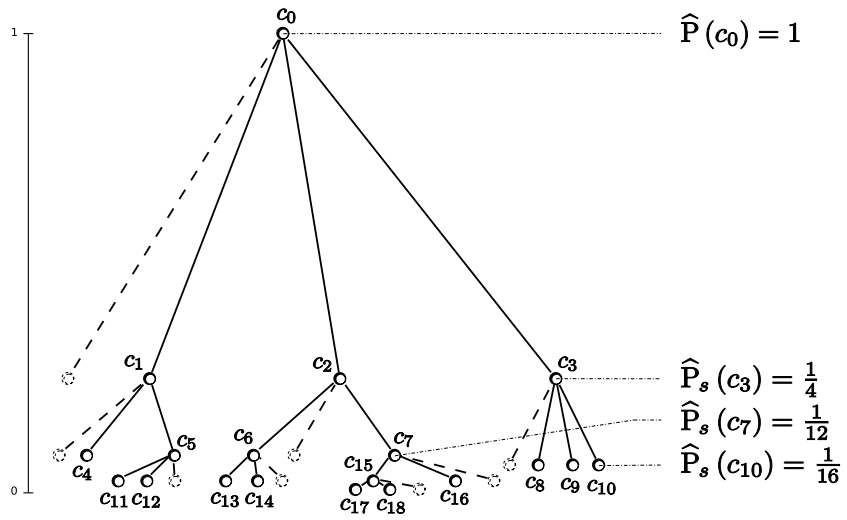


FIG. 4.6 – Application de l'approximation $\widehat{\mathrm{P}}_s$ avec $\widehat{\mathrm{P}}\left(c_0\right) = 1$ et $\epsilon = 1$

Nous remarquons que le relâchement de la complétude réduit l'influence du nombre de fils de chaque concept.

4.5.2 Approche ascendante

Approximation $\widehat{\mathrm{P}}_h$

La considération d'une distribution uniforme des instances pour les concepts de la même hauteur relève d'une approche ascendante dans laquelle à chaque niveau de hauteur[2], le nombre d'instances des concepts par rapport au nombre d'instances des concepts du niveau précédent est multiplié.

Si le contenu informationnel croît avec l'augmentation de la profondeur d'un concept c_i, il est tout aussi vrai qu'il décroît avec l'augmentation de sa hauteur

[2]un n$^{\text{ème}}$ niveau de hauteur contient tous les concepts de hauteur $n+1$. Les différents niveaux s'étalent donc du premier qui contient toutes les feuilles jusqu'au dernier contenant uniquement la racine.

h_i :

$$h_i = \begin{cases} 0 & , c_i \in c_0^{\propto} \\ 1 + \max\limits_{c_x \in c_i^{\succ}} h_x & , \text{sinon} \end{cases} \tag{4.13}$$

où $c_i^{\succ}$ est l'ensemble des fils de c_i.

Par hypothèse, toutes les feuilles ayant la même hauteur (hauteur nulle) ont le même nombre d'instances : $\forall c_i, c_j \in c_0^{\propto}, |\mathcal{E}_i| = |\mathcal{E}_j|$. Le cardinal de l'extension d'un concept noeud est issu de la multiplication du cardinal de l'extension de son fils le plus général par κ ($\kappa > 1$).

Dans ce cas, la probabilité qu'une instance soit associée à un concept c_i croît exponentiellement avec la hauteur de c_i dans $\mathcal{A}$. Un concept feuille a une probabilité minimale associée qui dépend de la hauteur de l'arbre et du nombre d'instances de la racine. Ainsi, on peut approximer $\mathrm{P}(c_i)$ (cf. figure 4.7 pour l'exemple) par :

$$\begin{aligned} \widehat{\mathrm{P}}_h(c_i) &= \begin{cases} \frac{\widehat{\mathrm{P}}(c_0)}{\kappa^{h_0}} & , c_i \in c_0^{\propto} \\ \kappa \cdot \max\limits_{c_x \in c_i^{\succ}} \widehat{\mathrm{P}}_h(c_x) & , \text{sinon} \end{cases} \\ &= \frac{\widehat{\mathrm{P}}(c_0)}{\kappa^{h_0 - h_i}} \end{aligned} \tag{4.14}$$

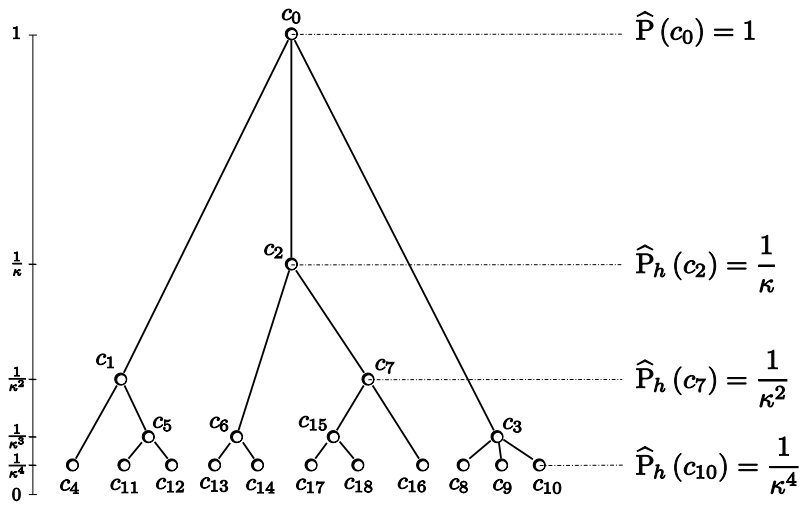


FIG. 4.7 – Application de l'approximation $\widehat{\mathrm{P}}_h$ avec $\widehat{\mathrm{P}}(c_0) = 1$

Si on prend le cas où la base du logarithme est égale à κ, le contenu informationnel d'un concept c_i est défini par :

$$\begin{aligned} \psi_h(c_i) &= -\log_\kappa \widehat{\mathrm{P}}_h(c_i) \\ &= -\log_\kappa \frac{\widehat{\mathrm{P}}(c_0)}{\kappa^{h_0 - h_i}} \\ &= h_0 - h_i - \log_\kappa \widehat{\mathrm{P}}(c_0) \\ &= h_0 - h_i + \psi(c_0) \end{aligned} \tag{4.15}$$

Approximation $\widehat{\mathrm{P}}_g$

Si on considère que la non spécialisation d'un concept feuille a un fondement on peut alors supposer une variabilité négligeable du nombre d'instances de chacun de ces concepts feuilles ; ceci en faisant l'hypothèse d'une équirépartition des instances de la racine sur les feuilles de l'arbre et ce de manière disjointe :

$$\forall c_i, c_j \in c_0^{\propto}, |\mathcal{E}_i| = |\mathcal{E}_j| \qquad \text{(uniformité des feuilles)}$$

$$\forall c_i, c_j \in c_0^{\propto}, \mathcal{E}_i \cap \mathcal{E}_j = \emptyset \qquad \text{(disjonction entre feuilles)}$$

où $c_0^{\propto}$ désigne l'ensemble des feuilles de la hiérarchie.

Lorsque la propriété de complétude est satisfaite, on obtient donc :

$$\begin{aligned} \widehat{\mathrm{P}}_g(c_i) &= \begin{cases} \frac{\widehat{\mathrm{P}}(c_0)}{|c_0^{\propto}|} & , c_i \in c_0^{\propto} \\ \sum_{c_x \in c_i^{\propto}} \widehat{\mathrm{P}}_g(c_x) & , \text{sinon} \end{cases} \\ &= \widehat{\mathrm{P}}(c_0) \cdot \frac{|c_i^{\propto}|}{|c_0^{\propto}|} \end{aligned} \tag{4.16}$$

où $c_i^{\propto}$ désigne l'ensemble des feuilles subsumées non strictement par c_i (si c_i est une feuille, $c_i^{\propto} = \{c_i\}$).

La figure 4.8 illustre le calcul de cette approximation. Par exemple, la probabilité attachée à c_2 se calcule en divisant le nombre de feuilles qu'il subsume par le nombre de feuilles total :

$$\widehat{\mathrm{P}}_g(c_2) = \frac{|\{c_{13}, c_{14}, c_{17}, c_{18}, c_{16}\}|}{|\{c_4, c_{11}, c_{12} c_{13}, c_{14}, c_{17}, c_{18}, c_{16}, c_8, c_9, c_{10}\}|} = \frac{5}{11}$$

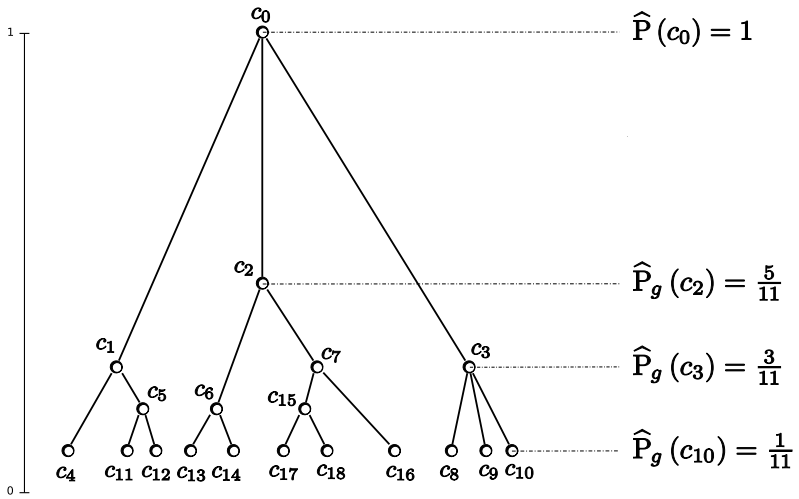


FIG. 4.8 – Application de l'approximation $\widehat{\mathrm{P}}_g$ avec $\widehat{\mathrm{P}}(c_0) = 1$

Le contenu informationnel $(-\log \widehat{\mathrm{P}}_g)$ issu de cette approximation rend compte du degré de généralité vis-à-vis des feuilles. La hauteur traduit une partie de l'information exploitée par ce degré de généralité. Il s'agit d'une approximation qui affine $\widehat{\mathrm{P}}_h$ en considérant en plus le nombre de fils du concept considéré et des concepts noeuds qu'il subsume.

Remarque 1. On peut redéfinir cette approximation pour éventuellement relâcher la contrainte de complétude. Dans ce cas, on ajoute ϵ concepts fils supplémentaires au niveau de chaque concept noeud. Si on considère ces concepts fils comme des feuilles, on redéfinit l'approximation précédente comme suit (cf. figure 4.9 pour l'exemple) :

$$\widehat{\mathrm{P}}_g(c_i) = \widehat{\mathrm{P}}(c_0) \cdot \frac{\sum_{c_x \in c_i^{\sqsupset}} \widehat{\omega}_g(c_x)}{\sum_{c_x \in \mathcal{C}} \widehat{\omega}_g(c_x)} \qquad \text{avec } \widehat{\omega}_g(c_i) = \begin{cases} 1 & \text{, si } c_i \in c_0^{\propto} \\ \epsilon & \text{, sinon} \end{cases} \tag{4.17}$$

où $c_i^{\sqsupset}$ est l'ensemble des concepts subsumés non strictement ($c_i \in c_i^{\sqsupset}$) par c_i.

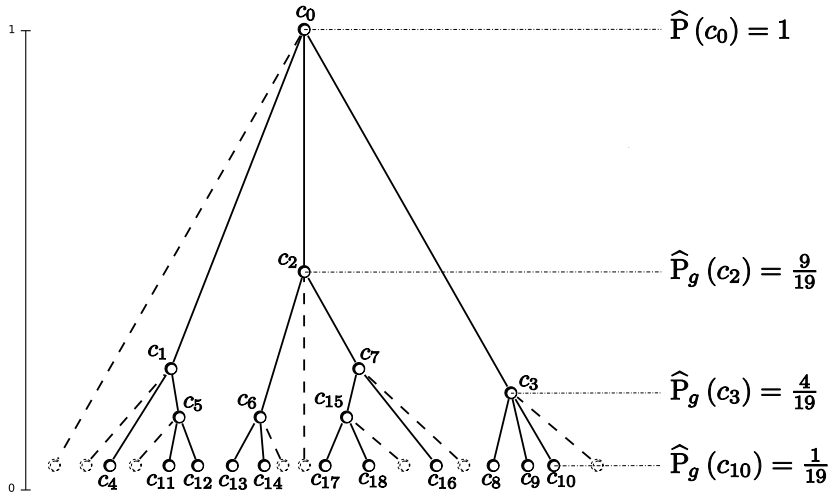


FIG. 4.9 – Application de l'approximation $\widehat{\mathrm{P}}_g$ avec $\widehat{\mathrm{P}}(c_0) = 1$ et $\epsilon = 1$

Proposition 4.1 *La redéfinition du contenu informationnel par Seco et al. [SVH04] rappelée dans l'introduction du paragraphe 4.5 correspond à une normalisation du contenu informationnel issu de* $\widehat{\mathrm{P}}_g$ *telle que* $\epsilon = 1$ *et* $\widehat{\mathrm{P}}(c_0) = 1$.

Preuve.

Considérons tout d'abord le contenu informationnel basé sur $\widehat{\mathrm{P}}_g$ avec $\epsilon = 1$ et $\widehat{\mathrm{P}}(c_0) = 1$:

$$\psi_g(c_i) = -\log_a\left(\frac{\left|c_i^{\sqsupset}\right|}{|\mathcal{C}|}\right) \tag{4.18}$$

Le contenu informationnel maximal $-\log_a\left(\frac{1}{|\mathcal{C}|}\right)$ est atteint pour chaque concept feuille et permet de normaliser $\psi_g(c_i)$ de manière à retrouver la proposition de Seco et al. :

$$\psi_{sec}(c_i) = \frac{\psi_g(c_i)}{-\log_a \frac{1}{|\mathcal{C}|}} \tag{4.19}$$

La normalisation qui est proposée par Seco et al. correspond à un changement d'échelle. En effet, leur choix a été de chercher à proposer une quantité d'information valuée sur l'intervalle $[0;1]$. La base du logarithme est le paramètre voué à cela. Dès lors, on s'aperçoit que l'on peut exprimer le contenu informationnel de Seco et al. (en utilisant la propriété $\log_u x = \frac{\log_v x}{\log_v u}$) en fonction de notre approximation $\widehat{\mathrm{P}}_g$:

$$\begin{aligned}
\psi_{sec}(c_i) &= \frac{\log_a\left(\frac{\left|c_i^{\sqsupset}\right|}{|\mathcal{C}|}\right)}{\log_a\left(\frac{1}{|\mathcal{C}|}\right)} \\
&= -\frac{\log_a\left(\frac{\left|c_i^{\sqsupset}\right|}{|\mathcal{C}|}\right)}{\log_a|\mathcal{C}|} \\
&= -\log_{|\mathcal{C}|}\frac{\left|c_i^{\sqsupset}\right|}{|\mathcal{C}|} \\
&= -\log_{|\mathcal{C}|}\widehat{\mathrm{P}}_g(c_i)
\end{aligned}$$

Remarque 2. Il est cependant plus pertinent de considérer que l'oubli d'un concept correspond à l'oubli d'un sous-arbre dont ce concept manquant est la racine. Nous pouvons supposer qu'un sous-arbre manquant est d'une importance comparable à celle des concepts frères du concept manquant. Nous redéfinissons l'approximation $\widehat{\mathrm{P}}_g$ en sachant qu'avec $\epsilon = 0$, on retrouve toujours un respect de la complétude (cf. figure 4.10 pour l'exemple) :

$$\begin{aligned}
\widehat{\mathrm{P}}_g(c_i) &= \widehat{\mathrm{P}}(c_0) \cdot \frac{\sum_{c_x \in c_i^{\sqsupset}} \widehat{\omega}(c_x)}{\sum_{c_x \in \mathcal{C}} \widehat{\omega}(c_x)} \\
\text{avec } \widehat{\omega}(c_i) &= \begin{cases} 1 & \text{, si } c_i \in c_0^{\propto} \\ \epsilon \cdot \sum_{c_x \in c_i^{\sqsupset}} \widehat{\omega}(c_x) & \\ \hline |c_i^{\succ}| & \text{, sinon} \end{cases}
\end{aligned} \tag{4.20}$$

4.5.3 Approche mixte

Les diverses approximations proposées exploitent différents aspects (e.g. profondeur de certains concepts, hauteur de la hiérarchie) de la structure hiérar-

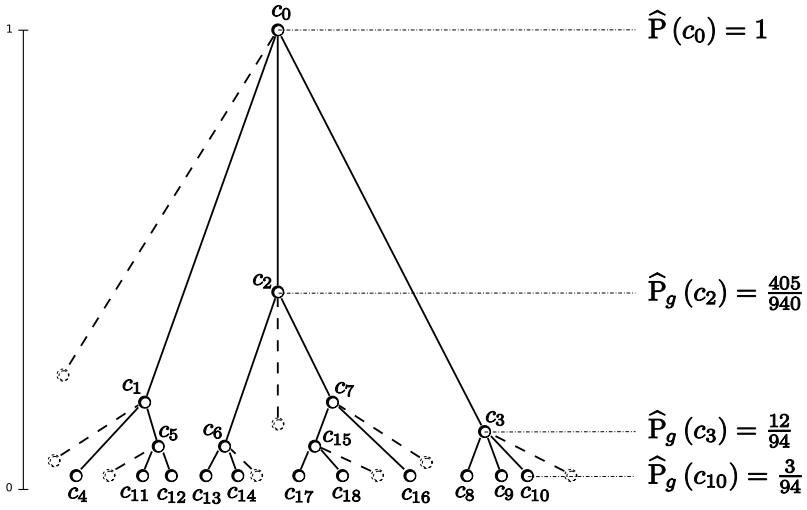


FIG. 4.10 – Application de l'approximation $\widehat{\mathrm{P}}_g$ avec $\widehat{\mathrm{P}}(c_0) = 1$ et $\epsilon = 1$

chique. Nous distinguons toutefois deux approches différentes : la méthode descendante qui est adoptée pour les approximations $\widehat{\mathrm{P}}_p$ et $\widehat{\mathrm{P}}_s$ et la méthode ascendante pour les approximations $\widehat{\mathrm{P}}_h$ et $\widehat{\mathrm{P}}_g$.

Ces deux approches comportent une part de réalisme selon l'ontologie considérée. Une alternative peut cependant s'avérer plus pertinente dans certains cas ; il s'agit de considérer à la fois l'influence du degré de spécificité vis-à-vis de la racine et celle du degré de généralité vis-à-vis des feuilles. Pour cela, nous proposons de combiner ces deux approches complémentaires en faisant la moyenne arithmétique de $\widehat{\mathrm{P}}_g$ et $\widehat{\mathrm{P}}_s$:

$$\widehat{\mathrm{P}}_{sg}(c_i) = \frac{\widehat{\mathrm{P}}_s(c_i) + \widehat{\mathrm{P}}_g(c_i)}{2} \tag{4.21}$$

Cette approximation est basée sur la moyenne arithmétique classique μ_1 (moyenne de Cauchy d'ordre 1). Nous avons fait ce choix pour que, lorsque la complétude est respectée par $\widehat{\mathrm{P}}_s$ et $\widehat{\mathrm{P}}_g$, $\widehat{\mathrm{P}}_{sg}$ la respecte également :

$$\begin{aligned}
&\widehat{\mathrm{P}}_{sg}(c_i) = \sum_{c_x \in c_i^{\propto}} \widehat{\mathrm{P}}_{sg}(c_x) \\
&\iff \mu_\alpha\left(\widehat{\mathrm{P}}_s(c_i), \widehat{\mathrm{P}}_g(c_i)\right) = \sum_{c_x \in c_i^{\propto}} \mu_\alpha(\widehat{\mathrm{P}}_s(c_x), \widehat{\mathrm{P}}_g(c_x)) \\
&\iff \mu_\alpha\left(\sum_{c_x \in c_i^{\propto}} \widehat{\mathrm{P}}_s(c_x), \sum_{c_x \in c_i^{\propto}} \widehat{\mathrm{P}}_g(c_x)\right) = \sum_{c_x \in c_i^{\propto}} \mu_\alpha\left(\widehat{\mathrm{P}}_s(c_x), \widehat{\mathrm{P}}_g(c_x)\right)
\end{aligned}$$

où μ_α désigne la moyenne de Cauchy d'ordre α.

Les approximations $\widehat{\mathrm{P}}_h$ et $\widehat{\mathrm{P}}_p$ exploitent moins finement la hiérachie de subsomption que les approximations $\widehat{\mathrm{P}}_g$ et $\widehat{\mathrm{P}}_s$ en réduisant la notion de généralité à

la hauteur et celle de spécificité à la profondeur. Cependant, elles sont toutes les deux aussi complémentaires et se prêtent également à être agrégées de la même manière que les approximations $\widehat{\mathrm{P}}_g$ et $\widehat{\mathrm{P}}_s$:

$$\widehat{\mathrm{P}}_{ph}\left(c_i\right) = \frac{\widehat{\mathrm{P}}_p\left(c_i\right) + \widehat{\mathrm{P}}_h\left(c_i\right)}{2} \tag{4.22}$$

Le choix d'une approximation doit être guidé par les connaissances disponibles, par les caractéristiques de l'arbre qui permettent une exploitation plus ou moins fine ainsi que par les objectifs applicatifs. En effet, plus l'information à notre disposition est conséquente (corpus, échantillon, etc.), plus la précision de la mesure sera grande. La forte variabilité du nombre de fils de chaque concept interviendra en faveur des approximations $\widehat{\mathrm{P}}_g$, $\widehat{\mathrm{P}}_s$ et $\widehat{\mathrm{P}}_{sg}$ en remplacement des approximations $\widehat{\mathrm{P}}_h$, $\widehat{\mathrm{P}}_p$ et $\widehat{\mathrm{P}}_{ph}$. Le choix d'une méthode ascendante, descendante ou mixte sera basé sur la pertinence de l'hypothèse sous-jacente. Enfin, d'éventuelles contraintes liées à l'objectif applicatif pourront également guider le choix de l'approximation.

4.6 Conclusion

Dans ce chapitre, nous avons repris le contenu informationnel d'un concept initialement introduit par Resnik afin de montrer qu'il correspond à une interprétation extensionnelle de l'arbre de subsomption. Tandis qu'une mesure de probabilité P définit l'interprétation extensionnelle dans le cas fictif où l'on dispose d'une information complète sur l'ensemble des concepts, de nombreuses approximations $\widehat{\mathrm{P}}$ sont envisageables en pratique. Nous avons proposé quelques approximations qui exploitent plus ou moins finement la structure hiérarchique sans recourir à un corpus.

Les approximations que nous proposons permettent une exploitation ciblée de l'information contenue dans la structure hiérarchique grâce à différentes hypothèses de distribution des instances sur les concepts de l'arbre. Notre réflexion a donnée lieu à des approximations qui relèvent principalement de deux approches duales :

- l'approche descendante
  - $\widehat{\mathrm{P}}_p$ repose sur l'hypothèse d'une réduction exponentielle de l'effectif des extensions des concepts avec l'augmentation de leur profondeur dans l'arbre :
  $$\widehat{\mathrm{P}}_p\left(c_i\right) = \frac{\widehat{\mathrm{P}}\left(c_0\right)}{\kappa^{p_i}}$$
  - $\widehat{\mathrm{P}}_s$ repose sur une hypothèse d'équirépartition des instances d'un concept père vers ses concepts fils, et ce depuis la racine jusqu'aux feuilles :
  $$\widehat{\mathrm{P}}_s\left(c_i\right) = \frac{\widehat{\mathrm{P}}_s\left(c_i^\star\right)}{\left|\left(c_i^\star\right)^\succ\right|}$$
- l'approche ascendante
  - $\widehat{\mathrm{P}}_h$ repose sur l'hypothèse d'un accroissement exponentiel de l'effectif des extensions des concepts avec l'augmentation de leur hauteur dans

l'arbre :

$$\widehat{\mathrm{P}}_h\left(c_i\right) = \frac{\widehat{\mathrm{P}}\left(c_0\right)}{\kappa^{h_0 - h_i}}$$

– $\widehat{\mathrm{P}}_g$ repose sur une hypothèse d'équirépartition des instances du concept racine sur les concepts feuilles et de regroupement des instances des concepts fils pour former l'extension de leur concept père depuis les feuilles jusqu'à la racine :

$$\widehat{\mathrm{P}}_g\left(c_i\right) = \widehat{\mathrm{P}}\left(c_0\right) \cdot \frac{|c_i^{\propto}|}{|c_0^{\propto}|}$$

Les approximations qui relèvent de l'approche ascendante considèrent que toutes les feuilles ont la même spécificité. De manière duale, l'approche descendante va au contraire les différencier le plus possible. C'est pourquoi nous avons proposé une approche mixte à travers l'agrégation d'approximations issues de ces deux approches complémentaires :

– l'approche mixte
  – $\widehat{\mathrm{P}}_{ph}$ agrège les approximations $\widehat{\mathrm{P}}_p$ et $\widehat{\mathrm{P}}_h$:

$$\widehat{\mathrm{P}}_{ph}\left(c_i\right) = \frac{\widehat{\mathrm{P}}_p\left(c_i\right) + \widehat{\mathrm{P}}_h\left(c_i\right)}{2}$$

  – $\widehat{\mathrm{P}}_{sg}$ agrège les approximations $\widehat{\mathrm{P}}_s$ et $\widehat{\mathrm{P}}_g$:

$$\widehat{\mathrm{P}}_{sg}\left(c_i\right) = \frac{\widehat{\mathrm{P}}_s\left(c_i\right) + \widehat{\mathrm{P}}_g\left(c_i\right)}{2}$$

Nous avons limité notre analyse à un arbre de subsomption mais notre approche se généralise à une hiérarchie de subsomption comme le montre le chapitre 6.

Analogie entre la manipulation de l'intension et de l'extension

5

Sommaire

Résumé

Il est difficile de rapprocher les mesures sémantiques qui considèrent la hiérarchie comme un graphe de celles qui utilisent le contenu informationnel. Plus largement, un cadre général pour la définition des mesures sémantiques est souhaitable pour une meilleure analyse et une plus grande maîtrise des mesures sémantiques. Pour initier ce chapitre, nous revenons sur la notion de contenu informationnel en montrant qu'elle fait le lien entre l'intension et l'extension d'un concept. Nous proposons ensuite un cadre fédérateur par le biais d'une analogie qui nous permet de réécrire les mesures sémantiques présentées au chapitre 2. Grâce à cette analogie,

nous adaptons à un arbre de subsomption les familles de similarités les plus utilisées présentées au chapitre 3. Nous étudions le comportement de ces familles de mesures dans des situations bien définies ainsi que leurs propriétés métriques et ordinales.

5.1 Introduction

Nous avons présenté au chapitre 2 les principales mesures sémantiques de la littérature destinées à l'évaluation de la ressemblance (ou absence de ressemblance) entre deux concepts d'une hiérarchie de subsomption (parfois restreinte à un arbre). Ces mesures généralement définies de façon ad hoc suivent pourtant des schémas bien connus (e.g coefficient de Jaccard [Jac01], coefficient de Dice [Dic45]).

En utilisant la notion de contenu informationnel et suivant l'approximation choisie, nous proposons l'ébauche d'un cadre fédérateur pour la définition de mesures sémantiques pour l'exploitation d'un arbre de subsomption. La plupart des mesures existantes trouvent leur place dans ce cadre formel qui permet d'une part d'analyser plus globalement leurs propriétés métriques et ordinales et d'autre part d'ouvrir la voie pour la définition de nombreuses mesures avec une signification maîtrisée.

Dans ce chapitre, nous limitons nos propositions à un arbre de subsomption. Celles-ci sont généralisées par la suite dans le chapitre 6.

5.2 Notion de contenu informationnel

5.2.1 Indicateur de spécificité

La relation de subsomption entre deux concepts indique que l'un d'eux est plus spécifique que l'autre. Il ne s'agit que d'une indication sur la spécificité d'un concept relativement à la spécificité d'un autre concept. De manière plus ambitieuse, le contenu informationnel d'un concept vise à évaluer la spécificité de ce concept. C'est en cela que l'approximation choisie nécessite des hypothèses complémentaires ou des informations externes à l'arbre de subsomption.

Suivant l'approximation choisie, le contenu informationnel peut indiquer une spécificité proportionnelle à la profondeur du concept considéré (avec $\widehat{\mathrm{P}}_p$), ou à la hauteur de l'arbre moins la hauteur du concept considéré (avec $\widehat{\mathrm{P}}_h$). L'intérêt du contenu informationnel est de pouvoir fournir une évaluation plus fine en considérant par exemple le nombre de fils de certains concepts ou encore des fréquences d'occurrences issues d'un corpus de textes (e.g. $\widehat{\mathrm{P}}_r$, $\widehat{\mathrm{P}}_s$, $\widehat{\mathrm{P}}_g$).

Quelque soit l'approximation choisie, l'échelle sur laquelle est valué le contenu informationnel est tout à fait arbitraire (cf. figure 5.1) et dépend de la base du logarithme utilisée. Nous nous intéressons maintenant à la définition de l'échelle la plus appropriée pour rendre le contenu informationnel intelligible.

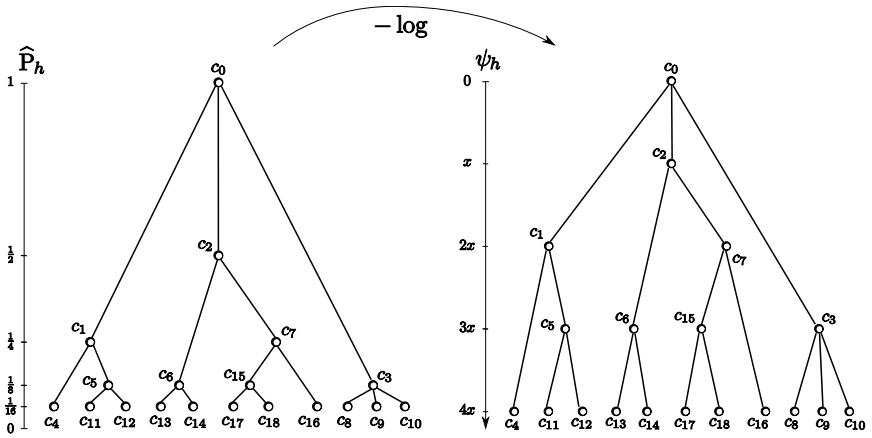


FIG. 5.1 – Passage de l'approximation $\widehat{\mathrm{P}}_h$ au contenu informationnel ψ_h

5.2.2 Choix de l'unité d'information

Si la profondeur est une approximation naturelle de la spécificité d'un concept, le contenu informationnel peut être considéré comme sa généralisation. Le contenu informationnel pourra non seulement être sensible à la profondeur, mais aussi aux densités locales, au nombre de feuilles subsumées ou à des éléments d'information externes comme un corpus.

L'influence de la base a du logarithme est telle que lorsque l'extension d'un concept est divisée par a, le contenu informationnel augmente d'une unité. Rendre le contenu informationnel intelligible, c'est faire le choix d'une échelle la plus intuitive pour l'interpréter. Devant la multiplicité et la diversité des utilisateurs potentiels, cette échelle n'est certainement pas unique. On peut par exemple vouloir considérer que l'unité d'information corresponde à la plus petite quantité d'information nécessaire à spécialiser un concept : $a = 2$. Si on considère un arbre donné, on peut fixer a au nombre de fils minimum d'un concept (sans considérer les feuilles) à condition que celui-ci soit supérieur à 2 ($a = max\{2; \min_{c_i \in \mathcal{C} - c_0^{\infty}} |c_i^{\succ}|\}$).

Il peut être intéressant de fixer la base du logarithme de manière à ce que la valeur obtenue pour chaque concept soit de l'ordre de sa profondeur. Il s'agit de faire en sorte que le contenu informationnel puisse être interprété comme une profondeur valuée sur $\mathbb{R}_+$ au lieu de $\mathbb{N}$. En d'autres termes, une unité d'information doit globalement représenter une profondeur d'une unité. Pour que l'unité d'information soit la même, il faut que la quantité totale d'information utilisée soit la même, c'est à dire que le contenu informationnel de l'ensemble des concepts $\mathcal{C}$ soit le même. Le calcul de la base du logarithme dépend donc de l'approximation utilisée. Pour ne considérer que la profondeur, il faut utiliser l'approximation $\widehat{\mathrm{P}}_p$ pour le calcul du contenu informationnel total de l'arbre de

subsomption avec $a = \kappa$:

$$\begin{aligned}\psi_p(\mathcal{C}) &= \psi(c_0) + \sum_{c_x \in \mathcal{C} - c_0} -\log_\kappa \frac{1}{\kappa} \\ &= \psi(c_0) + |\mathcal{C} - c_0| \\ &= \psi(c_0) + |\mathcal{C}| - 1\end{aligned}$$

Nous reprenons maintenant la formule générale du contenu informationnel de l'ensemble des concepts de l'arbre de subsomption de manière à poser la formule générale de la base du logarithme quelque soit l'approximation utilisée :

$$\begin{aligned}&\psi(\mathcal{C}) = \psi_p(\mathcal{C}) \\ &\iff \sum_{c_x \in \mathcal{C} - c_0} -\log_a \frac{\mathrm{P}(c_x)}{\mathrm{P}(c_x^\star)} = |\mathcal{C} - c_0| \\ &\iff a = \sqrt[|\mathcal{C}-\{c_0\}|]{\prod_{c_x \in \mathcal{C}-\{c_0\}} \frac{\mathrm{P}(c_x^\star)}{\mathrm{P}(c_x)}}\end{aligned}$$

On obtient finalement la moyenne géométrique des inverses des probabilités conditionnelles $P(c_x/c_x^\star)$. Si l'on prend par exemple le cas de l'approximation $\widehat{\mathrm{P}}_s$, cela correspond à la moyenne géométrique des nombres de fils du parent de chaque concept. Grâce à cette mise à l'échelle de la profondeur, fixer $\psi(c_i) = 1$ revient à estimer que le concept c_i est à une profondeur de 1. La signification du contenu informationnel est maintenant clairement intelligible quelque soit l'approximation utilisée.

5.3 Un cadre fédérateur pour un ensemble de mesures sémantiques

L'objectif de ce paragraphe est d'étudier les liens existants entre les similarités sémantiques les plus couramment utilisées dans la littérature (e.g. Rada, Resnik). En effet, bien que certaines ont été définies indépendamment des autres, nous montrons que de part leurs définitions, les différences peuvent être ténues. Ainsi, nous proposons un cadre formel qui permet de réécrire ces différentes mesures de façon à faciliter leur comparaison. Plus précisément, nous faisons l'analogie entre l'exploitation d'une description intensionnelle telle qu'elle est évoquée au chapitre 3 et l'exploitation du seul arbre de subsomption.

5.3.1 Analogie proposée

Nous avons vu dans le chapitre 4 que dans le cas où l'on ne dispose que d'un arbre de subsomption, nous pouvons considérer une interprétation extensionnelle particulière à l'aide d'une approximation de la probabilité associée à chaque concept. Quelque soit l'approximation choisie, le contenu informationnel évalue l'importance de l'intension d'un concept sur la base d'une approximation de l'effectif de son extension (cf. figure 5.2). On rappelle que l'intension

d'un concept c_i est noté $\mathcal{I}_i$ qui est un sous-ensemble de l'ensemble $\mathcal{I}$ des caractéristiques permettant de décrire tous les concepts de l'arbre de subsomption. L'extension d'un concept c_i qui est noté $\mathcal{E}_i$ est un sous-ensemble de l'ensemble $\mathcal{E}$ des instances du domaine considéré.

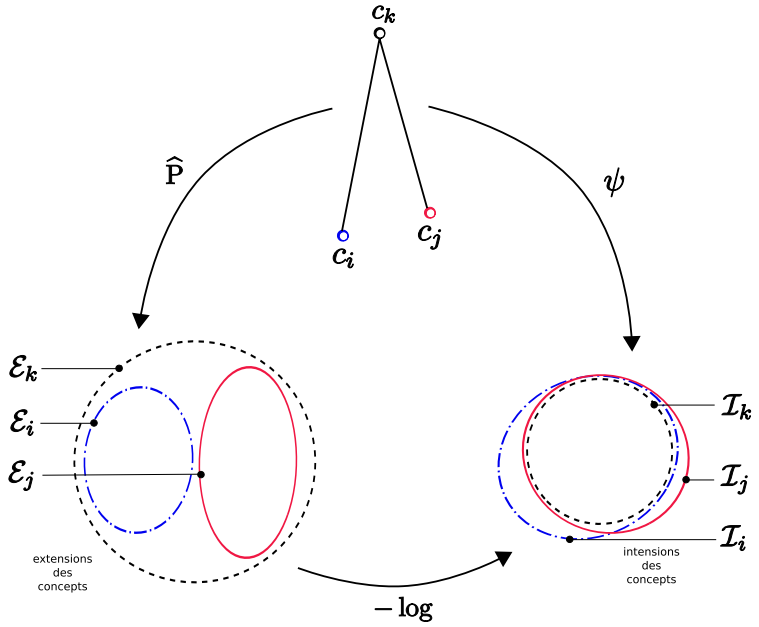


FIG. 5.2 – Principe du contenu informationnel

Les correspondances explicitées dans le tableau 5.2 traduisent l'analogie que l'on peut faire de manière à reprendre les travaux basés sur une représentation ensembliste pour tenter de les adapter à un arbre de subsomption. Étant donné deux concepts c_i et c_j, on rappelle que n_i, n_j et n_{ij} désignent respectivement l'importance des intensions $\mathcal{I}_i$, $\mathcal{I}_j$ et $\mathcal{I}_i \cap \mathcal{I}_j$. $\psi(c_i)$, $\psi(c_j)$ et $\psi(c_{ij})$ représentent respectivement le contenu informationnel de c_i, celui de c_j et le contenu informationnel partagé par c_i et c_j (contenu informationnel de leur subsumant le plus spécifique). Nous notons désormais $\psi^{\cap}(\{c_i, c_j\}) = \psi(c_{ij})$ le contenu informationnel partagé par c_i et c_j.

$(c_i, c_j) \in \mathcal{C}^2$	
n_i	$\psi(c_i)$
n_j	$\psi(c_j)$
n_{ij}	$\psi^{\cap}(\{c_i, c_j\})$

TAB. 5.2 – Analogie entre l'importance d'une description intensionnelle et le contenu informationnel

L'intérêt de cette analogie est double : nous allons pouvoir réécrire les mesures sémantiques existantes dans un même formalisme et utiliser des travaux sur les dissimilarités en taxonomie numérique en les adaptant aux arbres de subsomption.

5.3.2 Réécriture des mesures existantes

Nous montrons dans ce chapitre que la plupart des mesures sémantiques de la littérature qui exploitent uniquement un arbre de subsomption peuvent être considérées comme des adaptations de mesures usuelles définies sur une représentation ensembliste. Pour cela, il suffit de retrouver la mesure et l'approximation $\widehat{\mathrm{P}}$ adaptée (en précisant le contenu informationnel de la racine et la base du logarithme). Nous discutons dans le chapitre 6 des variations dans la prise en compte de l'héritage multiple.

Mesure de Resnik

La mesure de Resnik ne prend en compte que le contenu informationnel partagé par les deux concepts. Il s'agit donc d'une mesure analogue à la mesure de ressemblance de Restle (n_{ij}) avec l'approximation que propose Resnik ($\widehat{\mathrm{P}}_r$). On peut donc la reformuler :

$$res_{\mathcal{A}}(c_i, c_j) = \psi_r^{\cap}\left(\{c_i, c_j\}\right) \tag{5.1}$$

Précisons que la racine est considérée comme virtuelle ($\widehat{\mathrm{P}}\left(c_0\right) = 1$) et que la base du logarithme n'est pas explicitée par Resnik ($a = ?$).

Mesure de Jiang & Conrath

La mesure de Jiang & Conrath utilise également l'approximation de Resnik, mais évalue au contraire la quantité d'information qui différencie les deux concepts. Cette mesure est donc un équivalent de la dissemblance de Restle ($n_{i\bar{j}} + n_{\bar{i}j} = n_i + n_j - 2 \cdot n_{ij}$) avec l'approximation de Resnik :

$$jcn_{\mathcal{A}}(c_i, c_j) = \psi_r\left(c_i\right) + \psi_r\left(c_j\right) - 2 \cdot \psi_r^{\cap}\left(\{c_i, c_j\}\right) \tag{5.2}$$

Du fait de l'utilisation de l'approximation de Resnik, la racine est toujours considérée comme virtuelle ($\widehat{\mathrm{P}}\left(c_0\right) = 1$) et la base du logarithme non explicitée ($a = ?$).

Mesure de Rada

La mesure de Rada est également analogue à la dissemblance de Restle donc à la mesure de Jiang & conrath. La différence réside dans l'approximation qui n'utilise pas de corpus mais se base sur la profondeur avec $\widehat{\mathrm{P}}_p$:

$$rada_{\mathcal{A}}(c_i, c_j) = \psi_p\left(c_i\right) + \psi_p\left(c_j\right) - 2 \cdot \psi_p^{\cap}\left(\{c_i, c_j\}\right) \tag{5.3}$$

Précisons que Rada considère la longueur du chemin en nombre d'arcs. La racine est donc considérée comme virtuelle ($\widehat{\mathrm{P}}\left(c_0\right) = 1$) et la base du logarithme correspond à l'ordre de grandeur κ du nombre de fils d'un concept sur la globalité de l'arbre ($a = \kappa$).

La mesure de Leacock & Chodorow s'exprime comme nous l'avons vu au chapitre 3 en fonction de la mesure de Rada.

Mesure de Wu & Palmer

Les trois mesures précédentes ne tiennent compte que de l'information partagée ou de l'information qui différencie les deux concepts. La mesure de Wu & Palmer qui est analogue au coefficient de similarité de Dice $((2 \cdot n_{ij})/(n_i + n_j))$ prend en compte ces deux quantités. Elle repose sur l'utilisation de l'approximation $\widehat{\mathrm{P}}_p$:

$$wup_{\mathcal{A}}(c_i, c_j) = \frac{2 \cdot \psi_p^{\cap}(\{c_i, c_j\})}{\psi_p(c_i) + \psi_p(c_j)} \tag{5.4}$$

Notons que contrairement à ce que l'on pouvait observer dans les mesures précédentes (Resnik, Jiang & Conrath, Rada), la base du logarithme n'a pas d'effet dans la mesure de Wu & Palmer du fait de l'utilisation du rapport. Wu et Palmer considèrent la longueur des chemins en nombre de noeuds. cela revient à considérer la racine comme informative avec $\widehat{\mathrm{P}}(c_0) = \frac{1}{a}$ où a est la base du logarithme. Pour retrouver la variante introduite par Lin et Resnik, il faut considérer la racine comme virtuelle ($\widehat{\mathrm{P}}(c_0) = 1$).

Mesure de Lin

De la même manière que la mesure de Wu & Palmer, celle de Lin est analogue au coefficient de similarité de Dice mais cette fois-ci avec l'utilisation de l'approximation de Resnik :

$$lin_{\mathcal{A}}(c_i, c_j) = \frac{2 \cdot \psi_r^{\cap}(\{c_i, c_j\})}{\psi_r(c_i) + \psi_r(c_j)} \tag{5.5}$$

Du fait de l'utilisation de l'approximation de Resnik, la racine est considérée comme virtuelle ($\widehat{\mathrm{P}}(c_0) = 1$) tandis que la base du logarithme n'a pas d'effet du fait de la division.

Mesure de Stojanovic

La mesure de Stojanovic est assez proche de celle de Wu & Palmer puisqu'elle est analogue au coefficient de similarité de Jaccard $((n_{ij})/(n_i + n_j - n_{ij}))$ et utilise la même approximation $\widehat{\mathrm{P}}_p$ basée sur la profondeur :

$$sto_{\mathcal{A}}(c_i, c_j) = \frac{\psi_p^{\cap}(\{c_i, c_j\})}{\psi_p(c_i) + \psi_p(c_j) - \psi_p^{\cap}(\{c_i, c_j\})} \tag{5.6}$$

Tout comme pour la mesure de Wu & Palmer, la base du logarithme n'a pas d'effet et la racine est une racine informative avec $\widehat{\mathrm{P}}(c_0) = \frac{1}{a}$ où a est la base du logarithme.

Mesure de Zhong

Si le contenu informationnel ne permet pas de réécrire la mesure de Zhong, nous pouvons toutefois utiliser l'approximation $\widehat{\mathrm{P}}_p$ comme suit :

$$zhg_{\mathcal{A}}(c_i, c_j) = -\widehat{\mathrm{P}}_p(c_i) - \widehat{\mathrm{P}}_p(c_j) + 2 \cdot \widehat{\mathrm{P}}_p(c_{ij}) \tag{5.7}$$

La racine est considérée comme informative avec $\widehat{\mathrm{P}}(c_0) = \frac{1}{2}$. Cela montre que la seule différence avec Rada tient dans le fait que le log n'est pas utilisé. C'est le moyen mis en oeuvre par Zhong pour que deux concepts séparés par un même nombre d'arcs soient plus similaires lorsqu'ils sont plus profonds. Ce comportement obtenu ici de manière assez « artificielle » est naturellement intégré en tenant compte du contenu informationnel partagé par les deux concepts comme le font les mesures de Wu & Palmer, Stojanovic, Lin ou encore Resnik.

Proportion de Spécificité Partagée

Nous avons proposé dans une communication récente [BKHB06] une mesure, la proportion de spécificité partagée (psp). Elle est analogue au coefficient de Dice et utilise de façon inédite l'approximation $\widehat{\mathrm{P}}_s$ que nous avons présentée dans le chapitre 4 et étendue pour la prise en compte de l'héritage multiple dans le chapitre 6 :

$$psp_{\mathcal{A}}(c_i, c_j) = \frac{2 \cdot \psi_s^{\cap}(\{c_i, c_j\})}{\psi_s(c_i) + \psi_s(c_j)} \tag{5.8}$$

La base du logarithme n'a pas d'effet et nous avons laissé le choix dans notre proposition de considérer la racine soit comme informative ($\widehat{\mathrm{P}}(c_0) = \frac{1}{a}$ où a est la base du logarithme) soit comme virtuelle ($\widehat{\mathrm{P}}(c_0) = 1$).

5.3.3 Bilan

En plus de mettre à jour les liens que peuvent entretenir les principales mesures sémantiques de la littérature, l'analogie que nous proposons permet d'adapter l'ensemble des travaux présentés au chapitre 3. Comme nous l'avons vu, la ressemblance de Restle peut être adaptée pour retrouver la mesure de Resnik comme la dissemblance de Restle pour retrouver les mesures de Rada et Jiang & Conrath :

$$\widetilde{R}_{restle}(c_i, c_j) = \psi^{\cap}(\{c_i, c_j\}) \tag{5.9}$$

$$\widetilde{D}_{restle}(c_i, c_j) = \psi(c_i) + \psi(c_j) - 2 \cdot \psi^{\cap}(\{c_i, c_j\}) \tag{5.10}$$

Nous parlons de contenu informationnel différentiel noté $\psi^{\triangle}(\{c_i, c_j\})$ pour désigner la quantité d'information qui différencie les deux concepts $\psi^{\triangle}(\{c_i, c_j\}) = \psi(c_i) + \psi(c_j) - 2 \cdot \psi^{\cap}(\{c_i, c_j\})$. Tandis que l'adaptation du coefficient de Jaccard permet de retrouver la mesure de Stojanovic, celle de Dice est analogue aux mesures de Wu & Palmer et Lin. Ces deux coefficients font tous les deux partie de la famille de similarités σ_θ et Dice appartient également à la famille de similarités σ_α. Ces deux familles peuvent être adaptées comme suit :

$$\widetilde{\sigma}_\theta(c_i, c_j) = \left\{ \frac{\theta \cdot \psi^{\cap}(\{c_i, c_j\})}{(\theta - 2) \cdot \psi^{\cap}(\{c_i, c_j\}) + \psi(c_i) + \psi(c_j)} \right\}_{\theta \in \mathbb{R}_+^*} \tag{5.11}$$

$$\widetilde{\sigma}_\alpha(c_i, c_j) = \left\{ \frac{\psi^{\cap}(\{c_i, c_j\})}{\mu_\alpha\big(\psi(c_i), \psi(c_j)\big)} \right\}_{\alpha \in \mathbb{R}} \text{avec, } \mu_\alpha = \left(\frac{\psi(c_i)^\alpha + \psi(c_j)^\alpha}{2} \right)^{\frac{1}{\alpha}} \quad (5.12)$$

Nous ne reprenons pas de manière exhaustive l'ensemble des travaux évoqués au chapitre 3 qui peuvent être adaptés de cette manière. Il s'agit cependant d'une voix d'investigation parmi d'autres pour des applications qui ne se satisfont pas des mesures existantes. Il faut noter par exemple la possibilité d'adapter un certain nombre de mesures asymétriques que nous avons listées au chapitre 3.

5.4 Etude des familles $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$

Comme nous venons de le voir, les mesures sémantiques sont définies sur la base (1) de l'approximation utilisée pour le calcul du contenu informationnel des concepts et (2) de la définition de la mesure qui combine les contenus informationnels de différents concepts. Nous avons envisagé au chapitre précédent diverses approximations qui permettent de prendre en compte divers aspects (e.g. profondeur, hauteur) de l'arbre de subsomption. Nous proposons maintenant d'étudier les comportements et propriétés des similarités des familles $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$.

Nous cherchons d'une part à analyser les propriétés communes aux indices de chacune de ces familles et d'autre part à détecter les singularités des comportements de certains indices de la littérature qui se retrouvent dans ces familles.

Dans un premier temps, nous proposons de considérer des situations caractéristiques dans lesquelles un utilisateur aura une attente concernant le comportement de la mesure qu'il recherche. Nous faisons donc varier le contenu informationnel de deux concepts ainsi que leur contenu informationnel partagé et nous analysons l'impact sur les similarités des familles $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$. Dans un deuxième temps, nous nous attardons sur la préordonnance et le respect de l'inégalité triangulaire de Maguitman.

5.4.1 Comportement des similarités face aux variations du contenu informationnel

Nous considérons cinq cas différents qui nous permettent d'étudier le comportement des similarités sous l'influence des phénomènes suivants :

cas 1 l'augmentation du contenu informationnel que partagent deux concepts dont le contenu informationnel est fixe.

cas 2 l'augmentation du contenu informationnel que partagent deux concepts dont le contenu informationnel différenciel est fixe.

cas 3 l'augmentation du contenu informationnel différenciel de deux concepts dont le contenu informationnel partagé est fixe.

cas 4 la différence de spécificité entre deux concepts.

cas 5 la dilatation simultanée du contenu informationnel de chaque concept et de leur contenu informationnel partagé.

Nous analysons chacun des cas sur un exemple représentatif et nous nous sommes restreints à quelques indices de chaque famille ($\theta \in \{\frac{1}{2}, 1, 2, 3\}$, $\alpha \in \{-\infty, -1, 0, 1, 2, +\infty\}$) qui coïncident avec des indices que l'on retrouve dans la littérature.

Cas 1 (figure 5.3)

Sur la figure 5.3, les paramètres de l'étude de l'influence de l'augmentation du contenu informationnel que partagent deux concepts dont le contenu informationnel est fixe sont les suivants :

- $\psi(c_i) = \psi(c_j) = 12$
- $\psi^{\cap}(\{c_i, c_j\}) \in [0..12]$

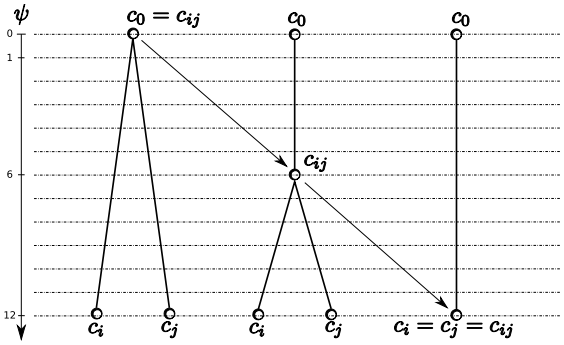


FIG. 5.3 – Influence de l'augmentation de $\psi^{\cap}(\{c_i, c_j\})$ avec $\psi(c_i)$ et $\psi(c_j)$ invariants

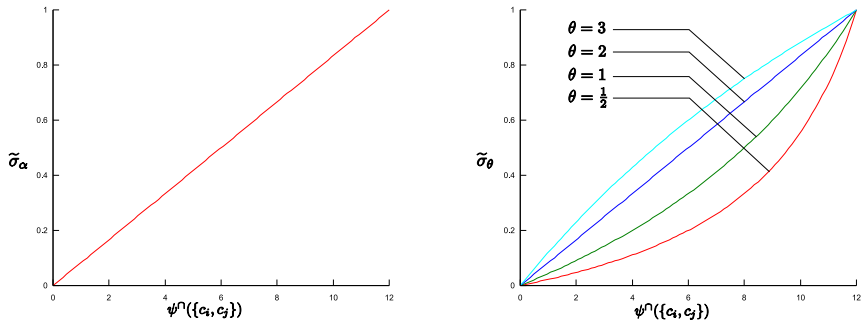
Toutes les similarités $\widetilde{\sigma}_\alpha$ réagissent exactement de la même manière avec une croissance linéaire d'une similarité nulle à une similarité maximale (cf. figure 5.4.A). Bien évidemment la similarité $\widetilde{\sigma}_{\theta=2}$ réagit comme les similarités $\widetilde{\sigma}_\alpha$ puisque $\widetilde{\sigma}_{\theta=2} = \widetilde{\sigma}_{\alpha=1}$. En revanche, on constate une croissance non linéaire convexe lorsque $\theta < 2$ et concave lorsque $\theta > 2$ (cf. figure 5.4.B). Ce cas souligne le rôle du paramètre θ qui fixe l'importance du contenu informationnel partagé vis-à-vis du contenu informationnel différentiel.

Cas 2 (figure 5.5)

Sur la figure 5.5, les paramètres de l'étude de l'influence de l'augmentation du contenu informationnel que partagent deux concepts dont le contenu informationnel différenciel est fixe sont les suivants :

- $\psi(c_i) = \psi(c_j) = \psi^{\cap}(\{c_i, c_j\}) + 3$
- $\psi^{\cap}(\{c_i, c_j\}) \in [0..9]$

Toutes les similarités $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$ réagissent exactement de la même manière avec une croissance non linéaire concave (cf. figure 5.6). On voit bien ici le rôle du contenu informationnel partagé qui influence la similarité de manière à ce que deux concepts séparés par un chemin de même longueur soient plus similaires lorsqu'ils sont plus spécifiques.



(A) (B)

FIG. 5.4 – Comportement de $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$ dans le cas 1 (figure 5.3)

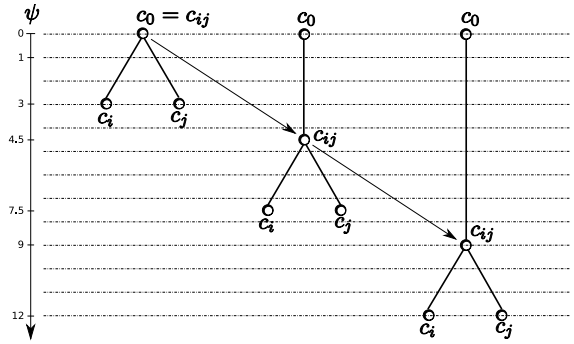


FIG. 5.5 – Influence de l'augmentation de $\psi^{\cap}(\{c_i, c_j\})$ avec $\psi^{\triangle}(\{c_i, c_j\})$ invariant

Cas 3 (figure 5.7)

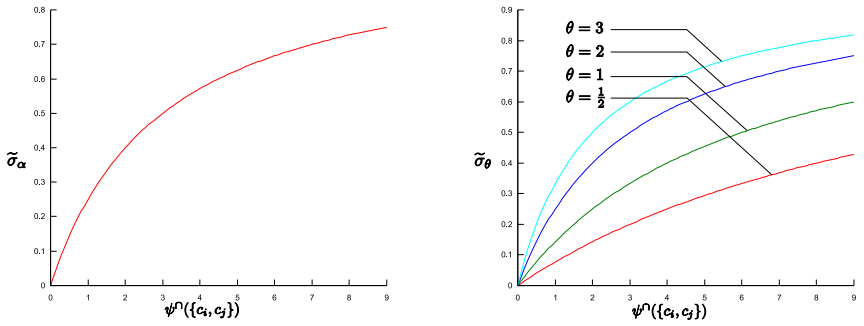
Sur la figure 5.7, les paramètres de l'étude de l'influence de l'augmentation du contenu informationnel différenciel de deux concepts dont le contenu informationnel partagé est fixe sont les suivants :

- $\psi^{\cap}(\{c_i, c_j\}) = 3$
- $\psi(c_i) = \psi(c_j)$
- $\psi(c_j) \in [3..12]$

Toutes les similarités $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$ ont un comportement similaire (cf. figure 5.8). On remarque en comparaison avec le test précédent, que les similarités $\widetilde{\sigma}_\alpha$ ont une vitesse de croissance en fonction de $\psi^{\cap}(\{c_i, c_j\})$ identique à leur vitesse de décroissance en fonction de $\psi^{\triangle}(\{c_i, c_j\})$.

Cas 4 (figure 5.9)

Sur la figure 5.9, les paramètres de l'étude de l'influence de la différence de spécificité entre les deux concepts sont les suivants :



(A) (B)

FIG. 5.6 – Comportement de $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$ dans le cas 2 (figure 5.5)

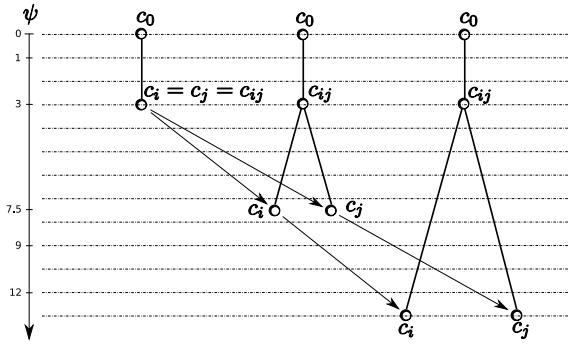
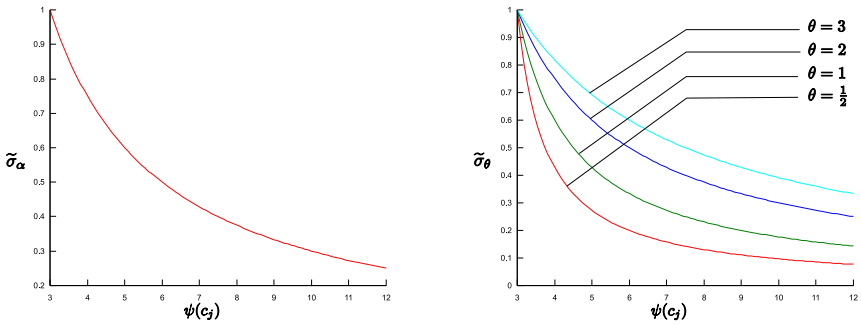


FIG. 5.7 – Influence de l'augmentation de $\psi^\triangle(\{c_i, c_j\})$ avec $\psi^\cap(\{c_i, c_j\})$ invariant

- $\psi^\cap(\{c_i, c_j\}) = 4$
- $\psi(c_i) = 16 - \psi(c_j)$
- $\psi(c_j) \in [8..12]$

La différence de spécificité entre les deux concepts n'influence pas les similarités $\widetilde{\sigma}_\theta$ qui sont constantes et prennent une valeur différente en fonction de l'importance qu'elles donnent à $\psi^\cap(\{c_i, c_j\})$ (cf. figure 5.10.B). Le choix de la moyenne utilisée dans une similarité de la famille $\widetilde{\sigma}_\alpha$ a pour effet de contrôler l'impact de la différence de spécificité sur la similarité comme le montre la figure 5.10.A. Le cas de la moyenne quadratique ($\alpha = 2$) est intéressant puisqu'elle décroît comme dans le cas extrême de l'utilisation de la fonction max ($\alpha = +\infty$). Si la différence de spécificité est considérée comme une différence supplémentaire entre les deux concepts, on pourra la pénaliser en utilisant une similarité $\widetilde{\sigma}_\alpha$ avec $\alpha > 1$.



(A) (B)

FIG. 5.8 – Comportement de $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$ dans le cas 3 (figure 5.7)

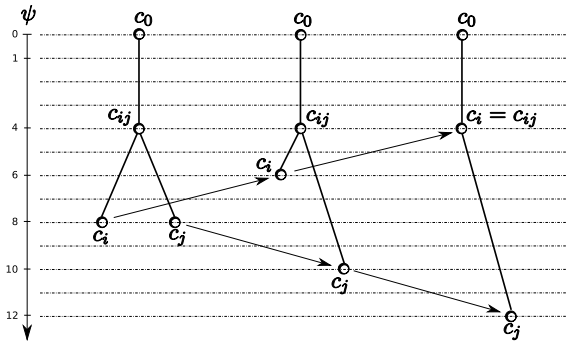


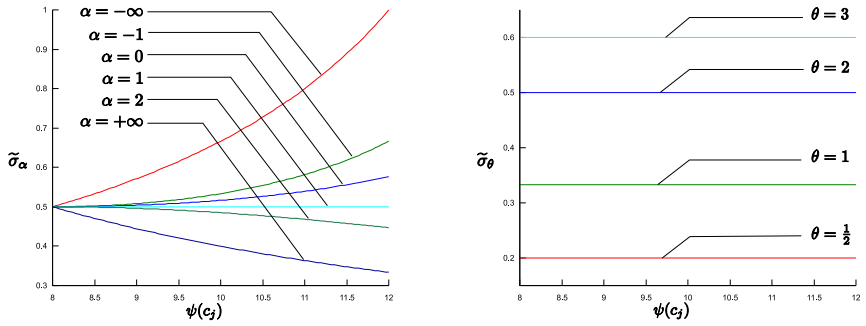
FIG. 5.9 – Influence de l'augmentation de $\psi(c_j) - \psi(c_i)$ avec $\psi^{\triangle}(\{c_i, c_j\})$ invariant

Cas 5 (figure 5.11)

Sur la figure 5.11, les paramètres de l'étude de l'influence de la dilatation simultanée du contenu informationnel de chaque concept et de leur contenu informationnel partagé sont les suivants :

- $\psi(c_i) = \psi(c_j) = \psi^{\cap}(\{c_i, c_j\}) + 3$
- $\psi^{\cap}(\{c_i, c_j\}) \in [0..9]$

Toutes les similarités $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$ sont constantes avec la dilatation simultanée du contenu informationnel de chaque concept et de leur contenu informationnel partagé (cf. figure 5.12). L'échelle du contenu informationnel étant fixée de manière arbitraire à l'aide du positionnement de la base du logarithme tandis que la similarité est normalisée sur $[0; 1]$, un autre comportement ne serait pas opportun. Autrement dit, un changement d'échelle du contenu informationnel n'a pas d'influence sur l'interprétation de la similarité entre deux concepts.



(A) (B)

FIG. 5.10 – Comportement de $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$ dans le cas 4 (figure 5.9)

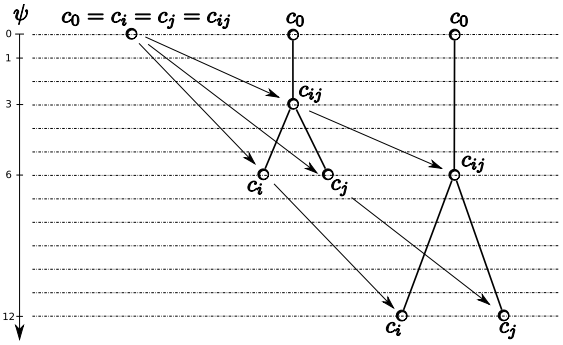


FIG. 5.11 – Influence de l'augmentation de $\psi(c_i)$, $\psi(c_j)$ et $\psi^{\cap}(\{c_i, c_j\})$

5.4.2 Propriétés métriques et ordinales

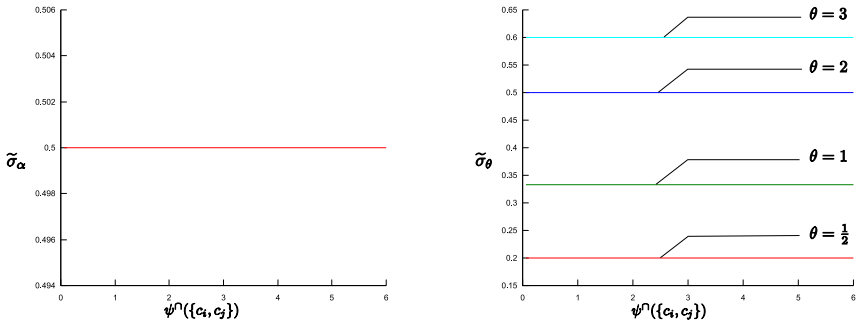
Dans les quelques travaux consacrés aux propriétés théoriques des mesures de similarité sémantiques (e.g. Lin [Lin98]), la majorité des auteurs se restreignent aux propriétés métriques. Dans cette lignée, nous analysons ainsi ici les propriétés métriques des indices $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$. Cependant, nous pensons qu'en pratique l'utilisateur se préoccupe plus souvent de l'ordre associé aux valeurs obtenues que des valeurs elles mêmes. En effet, ils ordonnent les paires de concepts selon les proximités quantifiées par ces mesures.

Définition 5.1 *Deux similarités sur $\mathcal{C}$ ont la même préordonnance si et seulement si*

$$\widetilde{\sigma}_1(c_i, c_j) \geq \widetilde{\sigma}_1(c_k, c_l) \iff \widetilde{\sigma}_2(c_i, c_j) \geq \widetilde{\sigma}_2(c_k, c_l) \tag{5.13}$$

Comme nous l'avons montré au cours du chapitre 3, l'inégalité triangulaire classique n'est pas adaptée à la signification des similarités normalisées comme $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$. Nous utilisons donc l'adaptation de l'inégalité triangulaire proposée par Maguitman [MMRV05] :

$$\widetilde{\sigma}(c_i, c_j) \geq \widetilde{\sigma}(c_i, c_k) \cdot \widetilde{\sigma}(c_k, c_j) \qquad \text{(inégalité de Maguitman)}$$



(A) (B)

FIG. 5.12 – Comportement de $\widetilde{\sigma}_\alpha$ et $\widetilde{\sigma}_\theta$ dans le cas 5 (figure 5.11)

Nous considérons tout d'abord la famille $\widetilde{\sigma}_\alpha$ puis la famille $\widetilde{\sigma}_\theta$.

Étude de la famille de similarités $\widetilde{\sigma}_\alpha$

Proposition 5.1 *Dans un arbre de subsomption, les similarités $\widetilde{\sigma}_\alpha$ n'ont pas la même préordonnance.*

Preuve.

Prenons un contre-exemple (cf. figure 5.13) avec quatre concepts : c_1, c_2, c_3, c_4 tel que c_1 subsume c_2, $\psi(c_1) = 1$, $\psi(c_2) = 3$, $\psi(c_3) = \psi(c_4) = 2$ et $\psi(c_{34}) = 1$.

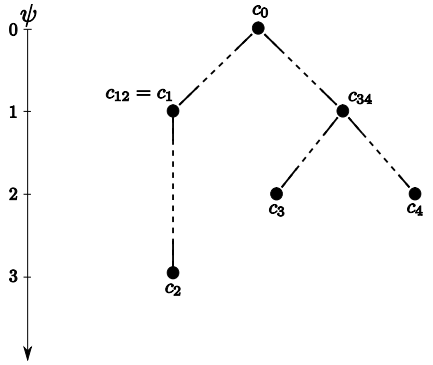


FIG. 5.13 – Contre-exemple concernant la préordonnance des similarités $\widetilde{\sigma}_\alpha$

On commence par calculer les moyennes de Cauchy sur cet exemple :

$$\mu_\alpha\big(\psi(c_1), \psi(c_2)\big) = \left(\frac{1+3^\alpha}{2}\right)$$

$$\mu_\alpha\big(\psi(c_3), \psi(c_4)\big) = 2$$

Si $\alpha > 1$ alors,

$$\begin{aligned} &\mu_\alpha\big(\psi\left(c_1\right), \psi\left(c_2\right)\big) > \mu_\alpha\big(\psi\left(c_3\right), \psi\left(c_4\right)\big) \\ \iff &\frac{\psi\left(c_1\right)}{\mu_\alpha\big(\psi\left(c_1\right), \psi\left(c_2\right)\big)} < \frac{\psi\left(c_{34}\right)}{\mu_\alpha\big(\psi\left(c_3\right), \psi\left(c_4\right)\big)} \end{aligned}$$

Si $\alpha = 1$ alors,

$$\begin{aligned} &\mu_\alpha\big(\psi\left(c_1\right), \psi\left(c_2\right)\big) = \mu_\alpha\big(\psi\left(c_3\right), \psi\left(c_4\right)\big) \\ \iff &\frac{\psi\left(c_1\right)}{\mu_\alpha\big(\psi\left(c_1\right), \psi\left(c_2\right)\big)} = \frac{\psi\left(c_{34}\right)}{\mu_\alpha\big(\psi\left(c_3\right), \psi\left(c_4\right)\big)} \end{aligned}$$

Si $\alpha < 1$ alors,

$$\begin{aligned} &\mu_\alpha\big(\psi\left(c_1\right), \psi\left(c_2\right)\big) < \mu_\alpha\big(\psi\left(c_3\right), \psi\left(c_4\right)\big) \\ \iff &\frac{\psi\left(c_1\right)}{\mu_\alpha\big(\psi\left(c_1\right), \psi\left(c_2\right)\big)} > \frac{\psi\left(c_{34}\right)}{\mu_\alpha\big(\psi\left(c_3\right), \psi\left(c_4\right)\big)} \end{aligned}$$

Donc la préordonnance des similarités $\widetilde{\sigma}_\alpha$ change suivant la valeur de α.

Proposition 5.2 *Dans un arbre de subsomption, si pour tout* $c_k \in c_i^{\sqsubseteq} - c_{ij}^{\sqsubset}$, $\widetilde{\sigma}_\alpha(c_i, c_k) \cdot \widetilde{\sigma}_\alpha(c_k, c_j) \leq \widetilde{\sigma}_\alpha(c_i, c_j)$ *alors pour tout* $c_x \in \mathcal{C}$, $\widetilde{\sigma}_\alpha(c_i, c_x) \cdot \widetilde{\sigma}_\alpha(c_x, c_j) \leq \widetilde{\sigma}_\alpha(c_i, c_j)$.

Preuve.

Si l'inégalité de Maguitman est respectée sur le chemin reliant c_i et le subsumant commun le plus spécifique de c_i et c_j, elle l'est également sur le chemin reliant c_j et le subsumant commun le plus spécifique de c_i et c_j et donc sur le chemin reliant c_i et c_j :

$$\begin{aligned} \forall c_k \in c_i^{\sqsubseteq} - c_{ij}^{\sqsubset}, \quad \widetilde{\sigma}_\alpha(c_i, c_k) \cdot \widetilde{\sigma}_\alpha(c_k, c_j) \leq \widetilde{\sigma}_\alpha(c_i, c_j) \\ \Longrightarrow \\ \forall c_k \in \left(c_i^{\sqsubseteq} \cup c_j^{\sqsubseteq}\right) - c_{ij}^{\sqsubset}, \quad \widetilde{\sigma}_\alpha(c_i, c_k) \cdot \widetilde{\sigma}_\alpha(c_k, c_j) \leq \widetilde{\sigma}_\alpha(c_i, c_j) \end{aligned}$$

Si on considère un concept c_x quelconque hors du chemin reliant c_i et c_j, il existe un concept c_k sur le chemin qui est plus similaire à c_i et c_j que ne l'est c_x :

$$\begin{aligned} \forall c_x \in \mathcal{C} - \left(\left(c_i^{\sqsubseteq} \cup c_j^{\sqsubseteq}\right) - c_{ij}^{\sqsubset}\right), \quad \exists c_k, \\ \widetilde{\sigma}_\alpha(c_i, c_k) \geq \widetilde{\sigma}_\alpha(c_i, c_x) \wedge \widetilde{\sigma}_\alpha(c_j, c_k) \geq \widetilde{\sigma}_\alpha(c_j, c_x) \end{aligned}$$

Lorsque l'inégalité de Maguitman est respectée sur le chemin reliant c_i et c_j, on peut en conclure que celle-ci est respectée sur l'ensemble des concepts de l'arbre de subsomption :

$$\forall c_x \in \mathcal{C}, \quad \widetilde{\sigma}_\alpha(c_i, c_x) \cdot \widetilde{\sigma}_\alpha(c_x, c_j) \leq \widetilde{\sigma}_\alpha(c_i, c_j)$$

Si l'inégalité de Maguitman est respectée avec l'ensemble des concepts sur le chemin reliant c_i et le subsumant commun le plus spécifique de c_i et c_j, elle l'est également avec tout autre concept.

Proposition 5.3 *Dans un arbre de subsomption, les similarités $\widetilde{\sigma}_\alpha$ ne respectent pas l'inégalité de Maguitman lorsque $\alpha < 0$ et respectent l'inégalité de Maguitman lorsque $\alpha \geq 0$.*

Preuve.

Nous considérons c_k appartenant au chemin reliant c_i et c_{ij}. En conséquence, $c_{ij} = c_{kj}$ et donc $\psi(c_{ij}) = \psi(c_{kj})$. Aussi, c_k subsume c_i ce qui implique que $c_{ik} = c_k$ et donc $\psi(c_{ik}) = \psi(c_k)$.

On définit alors les similarités nécessaires à la vérification de l'inégalité de Maguitman :

$$\widetilde{\sigma}_\alpha(c_i, c_k) = \frac{\psi(c_k)}{\mu_\alpha\big(\psi(c_i), \psi(c_k)\big)}$$
$$\widetilde{\sigma}_\alpha(c_k, c_j) = \frac{\psi(c_{kj})}{\mu_\alpha\big(\psi(c_k), \psi(c_j)\big)}$$
$$\widetilde{\sigma}_\alpha(c_i, c_j) = \frac{\psi(c_{kj})}{\mu_\alpha\big(\psi(c_i)), \psi(c_j)\big)}$$

L'inégalité de Maguitman est respectée si et seulement si :

$$\begin{aligned}
&\frac{\psi(c_k)}{\mu_\alpha\big(\psi(c_i), \psi(c_k)\big)} \cdot \frac{\psi(c_{kj})}{\mu_\alpha\big(\psi(c_k), \psi(c_j)\big)} \leq \frac{\psi(c_{kj})}{\mu_\alpha\big(\psi(c_i), \psi(c_j)\big)} \\
\iff &\frac{\psi(c_k)}{\mu_\alpha\big(\psi(c_i), \psi(c_k)\big)} \cdot \frac{\psi(c_{kj})}{\mu_\alpha\big(\psi(c_k), \psi(c_j)\big)} - \frac{\psi(c_{kj})}{\mu_\alpha\big(\psi(c_i), \psi(c_j)\big)} \leq 0 \\
\iff &\frac{\psi(c_k) \cdot \psi(c_{kj})}{\mu_\alpha\big(\psi(c_i), \psi(c_k)\big) \cdot \mu_\alpha\big(\psi(c_k), \psi(c_j)\big)} - \frac{\psi(c_k) \cdot \psi(c_{kj})}{\psi(c_k) \cdot \mu_\alpha\big(\psi(c_i), \psi(c_j)\big)} \leq 0 \\
\iff &\mu_\alpha\big(\psi(c_i), \psi(c_k)\big) \cdot \mu_\alpha\big(\psi(c_k), \psi(c_j)\big) - \psi(c_k) \cdot \mu_\alpha\big(\psi(c_i), \psi(c_j)\big) \geq 0
\end{aligned}$$

Or pour $\alpha = 0$,

$$\sqrt{\psi(c_i) \cdot \psi(c_k)} \cdot \sqrt{\psi(c_k) \cdot \psi(c_j)} - \psi(c_k) \cdot \sqrt{\psi(c_i) \cdot \psi(c_j)} = 0$$

De plus, si on fait évoluer α de 0 à $-\infty$, $\mu_\alpha\big(\psi(c_i), \psi(c_k)\big)$ et $\mu_\alpha\big(\psi(c_k), \psi(c_j)\big)$ décroissent plus vite que $\mu_\alpha\big(\psi(c_i), \psi(c_j)\big)$. D'où pour $\alpha < 0$,

$$\mu_\alpha\big(\psi(c_i), \psi(c_k)\big) \cdot \mu_\alpha\big(\psi(c_k), \psi(c_j)\big) - \psi(c_k) \cdot \mu_\alpha\big(\psi(c_i), \psi(c_j)\big) < 0$$

Donc l'inégalité de Maguitman n'est pas respectée lorsque c_k est sur le chemin entre c_i et le subsumant commun le plus spécifique de c_i et c_j. Dans un arbre de subsomption, les similarités $\widetilde{\sigma}_\alpha$ ne respectent pas l'inégalité de Maguitman lorsque $\alpha < 0$.

De plus, si on fait évoluer α de 0 à $+\infty$, $\mu_\alpha\big(\psi(c_i), \psi(c_k)\big)$ et $\mu_\alpha\big(\psi(c_k), \psi(c_j)\big)$ croissent plus vite que $\mu_\alpha\big(\psi(c_i), \psi(c_j)\big)$. D'où pour $\alpha \geq 0$,

$$\mu_\alpha\big(\psi(c_i), \psi(c_k)\big) \cdot \mu_\alpha\big(\psi(c_k), \psi(c_j)\big) - \psi(c_k) \cdot \mu_\alpha\big(\psi(c_i), \psi(c_j)\big) \geq 0$$

Donc l'inégalité de Maguitman est respectée lorsque c_k est sur le chemin entre c_i et le subsumant commun le plus spécifique de c_i et c_j. Comme démontré précédemment, si l'inégalité de Maguitman est respectée avec l'ensemble des concepts sur le chemin reliant c_i et le subsumant commun le plus spécifique de c_i et c_j, elle l'est également avec tout autre concept. Dans un arbre de subsomption, les similarités $\widetilde{\sigma}_\alpha$ respectent l'inégalité de Maguitman lorsque $\alpha \geq 0$.

Étude de la famille de similarités $\widetilde{\sigma}_\theta$

Proposition 5.4 *Dans un arbre de subsomption, les similarités $\widetilde{\sigma}_\theta$ ont la même préordonnance.*

Preuve.

Si $(\psi(c_i)) + (\psi(c_j)) - 2 \cdot (\psi(c_{ij})) = 0$ alors,

$$\forall \theta \in \mathbb{R}_+^*, \widetilde{\sigma}_1(c_i, c_j) = \widetilde{\sigma}_\theta(c_i, c_j)$$

Si $(\psi(c_i)) + (\psi(c_j)) - 2 \cdot (\psi(c_{ij})) \neq 0$ on pose :

$$\begin{aligned}
\frac{\widetilde{\sigma}_1(c_i, c_j)}{\widetilde{\sigma}_\theta(c_i, c_j)} &= \frac{\psi(c_{ij}) \cdot [\psi(c_i) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{ij})]}{[\psi(c_i) + \psi(c_j) - \psi(c_{ij})] \cdot \theta \cdot \psi(c_{ij})} \\
&= \frac{\psi(c_i) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{ij})}{\theta \cdot [\psi(c_i) + \psi(c_j) - \psi(c_{ij})]} \\
&= \frac{1}{\theta} \cdot \frac{\psi(c_i) + \psi(c_j) - \psi(c_{ij}) + (\theta - 1) \cdot \psi(c_{ij})}{\psi(c_i) + \psi(c_j) - \psi(c_{ij})} \\
&= \frac{1 + (\theta - 1) \cdot \widetilde{\sigma}_1(c_i, c_j)}{\theta}
\end{aligned}$$

Comme $\widetilde{\sigma}_1 \in [0; 1]$ et $\theta > 0$, $\frac{1+(\theta-1)\cdot\widetilde{\sigma}_1(c_i,c_j)}{\theta} \neq 0$ et

$$\widetilde{\sigma}_\theta(c_i, c_j) = \frac{\theta \cdot \widetilde{\sigma}_1(c_i, c_j)}{1 + (\theta - 1) \cdot \widetilde{\sigma}_1(c_i, c_j)}$$

Ainsi,

$$\forall (c_i, c_j, c_k, c_l) \in \mathcal{C}^4, \widetilde{\sigma}_1(c_i, c_j) \geq \widetilde{\sigma}_1(c_k, c_l) \iff \widetilde{\sigma}_\theta(c_i, c_j) \geq \widetilde{\sigma}_\theta(c_k, c_l)$$

Les $\widetilde{\sigma}_\theta(c_i, c_j)$ ont donc tous la même préordonnance.

Proposition 5.5 *Dans un arbre de subsomption, les similarités $\widetilde{\sigma}_\theta$ respectent l'inégalité de Maguitman lorsque $\theta \leq 1$ et ne respectent pas l'inégalité de Maguitman lorsque $\theta > 1$.*

Preuve.

Nous considérons c_k appartenant au chemin reliant c_i et c_{ij}. En conséquence, $c_{ij} = c_{kj}$ et donc $\psi(c_{ij}) = \psi(c_{kj})$. Aussi, c_k subsume c_i ce qui implique que, $c_{ik} = c_k$ et donc $\psi(c_{ik}) = \psi(c_k)$.

On définit alors les similarités nécessaires à la vérification de l'inégalité de Maguitman :

$$\widetilde{\sigma}_\theta(c_i, c_k) = \frac{\theta \cdot (\psi(c_k))}{(\psi(c_i)) + (\theta - 1) \cdot (\psi(c_k))}$$
$$\widetilde{\sigma}_\theta(c_k, c_j) = \frac{\theta \cdot (\psi(c_{kj}))}{(\psi(c_k)) + (\psi(c_j)) + (\theta - 2) \cdot (\psi(c_{kj}))}$$
$$\widetilde{\sigma}_\theta(c_i, c_j) = \frac{\theta \cdot (\psi(c_{kj}))}{(\psi(c_i)) + (\psi(c_j)) + (\theta - 2) \cdot (\psi(c_{kj}))}$$

L'inégalité de Maguitman est respectée si et seulement si :

$$\frac{\theta \cdot \psi(c_k)}{\psi(c_i) + (\theta - 1) \cdot \psi(c_k)} \cdot \frac{\theta \cdot \psi(c_{kj})}{\psi(c_k) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{kj})} \leq \frac{\theta \cdot \psi(c_{kj})}{\psi(c_i) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{kj})}$$

$$\iff \frac{\theta \cdot \psi(c_k)}{\psi(c_i) + (\theta - 1) \cdot \psi(c_k)} \cdot \frac{\theta \cdot \psi(c_{kj})}{\psi(c_k) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{kj})} - \frac{\theta \cdot \psi(c_{kj})}{\psi(c_i) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{kj})} \leq 0$$

$$\iff \frac{\theta^2 \cdot \psi(c_k) \cdot \psi(c_{kj})}{\Big(\psi(c_i) + (\theta - 1) \cdot \psi(c_k)\Big) \cdot \Big(\psi(c_k) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{kj})\Big)} - \frac{\theta^2 \cdot \psi(c_k) \cdot \psi(c_{kj})}{\Big(\theta \cdot \psi(c_k)\Big) \cdot \Big(\psi(c_i) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{kj})\Big)} \leq 0$$

$$\iff \Big(\psi(c_i) + (\theta - 1) \cdot \psi(c_k)\Big) \cdot \Big(\psi(c_k) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{kj})\Big) - \Big(\theta \cdot \psi(c_k)\Big) \cdot \Big(\psi(c_i) + \psi(c_j) + (\theta - 2) \cdot \psi(c_{kj})\Big) \geq 0$$

$$\begin{aligned}\iff \Big(&\psi(c_i) \cdot \psi(c_k) + \psi(c_i) \cdot \psi(c_j) \\ &+ (\theta - 2) \cdot \psi(c_i) \cdot \psi(c_{kj}) + (\theta - 1) \cdot \psi(c_k)^2 \\ &+ (\theta - 1) \cdot \psi(c_k) \cdot \psi(c_j) + (\theta - 1) \cdot (\theta - 2) \cdot \psi(c_k) \cdot \psi(c_{kj})\Big) \\ &- \Big(\theta \cdot \psi(c_i) \cdot \psi(c_k) + \theta \cdot \psi(c_k) \cdot \psi(c_j) \\ &+ \theta \cdot (\theta - 2) \cdot \psi(c_k) \cdot \psi(c_{kj})\Big) \geq 0\end{aligned}$$

$$\begin{aligned}\iff &\psi(c_i) \cdot \psi(c_j) + (\theta - 2) \cdot \psi(c_i) \cdot \psi(c_{kj}) \\ &+ (\theta - 1) \cdot \psi(c_k)^2 - (\theta - 1) \cdot \psi(c_i) \cdot \psi(c_k) \\ &- \psi(c_k) \cdot \psi(c_j) - (\theta - 2) \cdot \psi(c_k) \cdot \psi(c_{kj}) \geq 0\end{aligned}$$

$$\begin{aligned}
\iff & \psi(c_j) \cdot \Big(\psi(c_i) - \psi(c_k)\Big) \\
& \quad + \Big((\theta - 2) \cdot \psi(c_{kj})\Big) \cdot \Big(\psi(c_i) - \psi(c_k)\Big) \\
& \quad + \Big((\theta - 1) \cdot \psi(c_k)\Big) \cdot \Big(\psi(c_k) - \psi(c_i)\Big) \geq 0
\end{aligned}$$

$$\iff \Big(\psi(c_i) - \psi(c_k)\Big) \cdot \Big((\theta - 2) \cdot \psi(c_{kj}) + (\theta - 1) \cdot \psi(c_k) + \psi(c_j)\Big) \geq 0$$

$$\begin{aligned}
\iff & \Big(\psi(c_i) - \psi(c_k)\Big) \cdot \Bigg((1 - \theta) \cdot \Big(\psi(c_k) - \psi(c_{kj})\Big) \\
& \quad + \Big(\psi(c_j) - \psi(c_{kj})\Big)\Bigg) \geq 0
\end{aligned}$$

Lorsque $0 < \theta \leq 1$, cette inégalité est vérifiée et donc l'inégalité de Maguitman est respectée. Au contraire, lorsque $\theta > 1$, il suffit que $\psi(c_j) - \psi(c_{kj}) = 0$ pour que cette inégalité et donc l'inégalité de Maguitman ne soit pas vérifiée.

5.5 Conclusion

Le concept de similarité est fondamental dans beaucoup de domaines (e.g. classification, IA, psychologie, ...). A l'origine, les mesures sont souvent construites pour atteindre des objectifs précis. Cependant, quelques mesures (e.g. Jaccard [Jac01], Dice [Dic45]) ont montré leur pertinence dans des applications très diverses. Aujourd'hui les similarités suscitent un regain d'intérêt du fait du succès des ontologies en Ingénierie des Connaissances.

Nous avons montré que les mesures sémantiques de la littérature suivent généralement des schémas classiques (e.g. coefficient de Jaccard, coefficient de Dice). Dans cet esprit, nous avons mis en évidence une analogie avec les mesures ensemblistes en utilisant la notion de contenu informationnel. Cette analogie contribue à préciser la signification des mesures sémantiques de la littérature mais permet également d'adapter un certain nombre de travaux présentés au chapitre 3 pour exploiter un arbre de subsomption. Elle offre également une voie d'investigation pour la définition de mesures sémantiques asymétriques. En effet, les travaux sur la qualité des règles en ECD peuvent potentiellement être adaptés.

Nous avons mis en évidence la possibilité de retrouver les mesures sémantiques de la littérature et plus généralement de définir une mesure sémantique en choisissant :

- une mesure ensembliste adaptée par le biais de notre analogie
- une approximation $\widehat{\mathrm{P}}$ sur laquelle repose le contenu informationnel
- le statut de la racine (informative ou virtuelle)
- la base du logarithme a nécessaire au contenu informationnel (lorsque celle-ci a un effet sur la mesure)

Les mesures les plus utlisées dans tous les domaines font partie des familles de similarités σ_α et σ_θ dont nous avons étudié le comportement dans certaines configurations précises d'un arbre de subsomption. Cette analyse permet de guider le choix d'une mesure en fonction du comportement attendu. Nous avons également étudié la préordonnance et le respect de l'inégalité de Maguitman (adaptation de l'inégalité triangulaire) [MMRV05] pour ces familles de similarités.

Généralisation à l'héritage multiple

6

Sommaire

Résumé

L'héritage multiple est couramment utilisé lors de la construction d'une hiérarchie de subsomption. Cependant, dans les chapitre 4 et 5, nous avons limité nos propositions à un arbre de subsomption. Dans ce chapitre, nous approfondissons la notion de contenu informationnel pour l'adapter à une hiérarchie de subsomption. Pour cela, nous revenons brièvement sur la notion d'interprétation extensionnelle avant de proposer une généralisation du contenu informationnel à un sous-ensemble de concepts. Nous reprenons également la notion de contenu informationnel partagé qui nous permet notamment de reformuler le contenu informationnel d'un ensemble

de concepts pour en simplifier l'implémentation. Nous reprenons les diverses approximations du chapitre 4 pour les adapter lorsque cela est nécessaire à une hiérarchie de subsomption. Nous reprenons notre analogie pour tenir compte de l'héritage multiple et discutons de son impact dans la réécriture des mesures de la littérature et sur leurs propriétés. En dernier lieu, nous généralisons notre approche au calcul de la similarité entre deux sous-ensembles de concepts.

6.1 Introduction

L'héritage multiple pose des problèmes spécifiques qui nous amènent à approfondir nos réflexions du chapitre 4. Cette généralisation a tout d'abord une incidence sur le calcul de certaines approximations et notamment au niveau des concepts ayant plusieurs parents. Quelque soit l'approximation utilisée, le calcul de la quantité d'information commune à deux concepts peut faire intervenir le contenu informationnel non plus d'un seul concept, mais d'un sous-ensemble de concepts.

Dans un premier temps, nous rediscutons la notion d'interpétation extensionnelle après avoir complété les notations introduites au chapitre 4. Nous proposons ensuite les notions de contenu informationnel global et partagé pour un sous-ensemble de concepts de $\mathcal{C}$. Nous présentons également une extension des approximations du chapitre 4 pour l'exploitation d'une hiérarchie de subsomption. L'analogie que nous avons proposé au chapitre 5 fait l'objet d'une génaralisation pour l'héritage multiple.

6.2 Notations

Nous réutilisons les relations définies au chapitre 4 auxquelles nous ajoutons certaines relations du type $\mathcal{R}!$ et $\mathcal{R}!$ avec $\mathcal{R} \in \{\sqsupseteq, \sqsupset, \sqsubseteq, \sqsubset, \prec, \succ, \propto\}$ tel que $\mathcal{R}!$ et $\mathcal{R}!$ sont des sous-ensembles disjoints de $\mathcal{R}$:

$$c_i \; \mathcal{R}! \; c_j \iff [c_i \; \mathcal{R} \; c_j \wedge \exists! c_x (c_j \prec c_x)]$$
$\hookrightarrow c_j$ est une image de c_i par $\mathcal{R}$ ayant un et un seul père.

$$c_i \; \mathcal{R}! \; c_j \iff [c_i \; \mathcal{R} \; c_j \wedge \exists c_x, c_y (c_x \neq c_y \wedge c_j \prec c_x \wedge c_j \prec c_y)]$$
$\hookrightarrow c_j$ est une image de c_i par $\mathcal{R}$ ayant plusieurs parents (au moins deux).

Nous introduisons également deux simplifications d'écriture $\cup_{\mathcal{C}_i}^{\mathcal{R}}$ et $\cap_{\mathcal{C}_i}^{\mathcal{R}}$ permettant une manipulation plus concise d'un ensemble de concepts $\mathcal{C}_i$:

$$\cup_{\mathcal{C}_i}^{\mathcal{R}} = \bigcup_{c_x \in \mathcal{C}_i} c_x^{\mathcal{R}}$$
$$\cap_{\mathcal{C}_i}^{\mathcal{R}} = \bigcap_{c_x \in \mathcal{C}_i} c_x^{\mathcal{R}}$$

6.3 Notion de contenu informationnel

6.3.1 Interprétation extensionnelle d'une hiérarchie de subsomption

Dans une hiérarchie de subsomption, l'extension d'un concept ayant plusieurs parents (successeurs) est incluse dans l'intersection des extensions de ses parents. Nous reprenons l'exemple du chapitre 4 auquel nous ajoutons un lien de subsomption entre c_5 et c_{13}. Une interpétation extensionnelle de cet exemple de hiérarchie de subsomption (qui est repris par la suite) est illustrée par la figure 6.1.

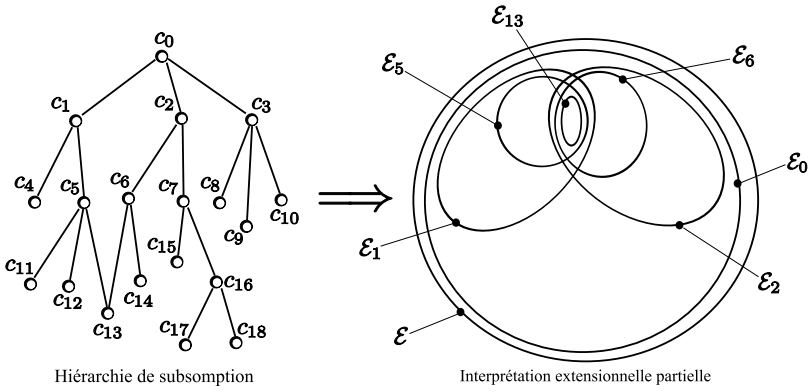


FIG. 6.1 – Interprétation extensionnelle partielle d'une hiérarchie de subsomption

La hiérarchie du fait de l'héritage multiple au niveau de c_{13} impose une intersection entre $\mathcal{E}_5$ et $\mathcal{E}_6$ ou encore $\mathcal{E}_1$ et $\mathcal{E}_2$. On peut être amené du fait d'une information externe qui le précise ou bien par hypothèse, à considérer une disjonction lorsque celle-ci n'est pas explicitement contredite par la structure hiérarchique :

$$\forall c_i, c_j \in \mathcal{C}, c_i^{\sqsupseteq} \cap c_j^{\sqsupseteq} = \emptyset \Longrightarrow \mathcal{E}_i \cap \mathcal{E}_j = \emptyset \qquad \text{(disjonction)}$$

6.3.2 Contenu informationnel global

L'exploitation d'une hiérarchie de subsomption nécessite le calcul du contenu informationnel d'un sous-ensemble de concepts que nous appelons contenu informationnel global. C'est pourquoi nous proposons maintenant de généraliser le contenu informationnel à un sous-ensemble de concepts $\mathcal{C}_i \subseteq \mathcal{C}$. Nous rappelons que le contenu informationnel $\psi(c_i)$ d'un concept c_i peut être interprété comme la quantité d'information apportée par l'événement « l'instance appartient à l'extension du concept c_i ». Nous définissons le contenu informationnel global $\psi^{\cup}(\mathcal{C}_i)$ d'un sous-ensemble de concepts $\mathcal{C}_i$ comme la quantité d'information apporté par au moins un des événements du type « l'instance appartient à l'extension du concept c_x » avec c_x appartenant à $\mathcal{C}_i$. Toutefois, la quantité

d'information globale d'un ensemble $\mathcal{C}_i$ de concepts ne se résume pas à la somme des quantités d'information de chaque concept de l'ensemble $\mathcal{C}_i$ dans la mesure où une partie de l'information nécessaire à la description d'un concept peut être incluse dans l'information nécessaire à la description d'un autre concept. Par souci de clarté, nous nous restreignons tout d'abord à un arbre avant de généraliser notre proposition.

Cas d'un arbre

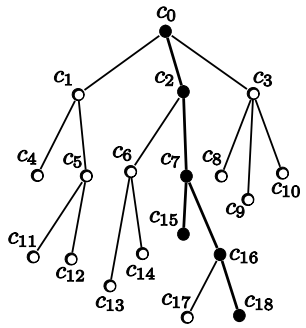
Nous pouvons tout d'abord remarquer que lorsqu'un concept c_i est subsumé strictement par un concept c_j ($c_i \sqsubset c_j$) alors $\mathrm{P}(c_i) < \mathrm{P}(c_j)$ et $\mathrm{P}(c_i) = \mathrm{P}(c_j)\,\rho_{ij}$ avec ρ_{ij} la probabilité qu'une instance quelconque appartienne à l'extension du concept c_i sachant qu'elle appartient à celle du concept c_j (en effet, $\Pr(\mathcal{E}_i/\mathcal{E}_j) = \frac{\Pr(\mathcal{E}_i \cap \mathcal{E}_j)}{\Pr(\mathcal{E}_j)} = \frac{\Pr(\mathcal{E}_i)}{\Pr(\mathcal{E}_j)}$). Par suite, le lien entre le contenu informationnel de c_i et celui de c_j est caractérisé par l'équivalence $\psi(c_i) = \psi(c_j) - \log_a \rho_{ij}$. Le concept c_i apporte la quantité d'information $-\log_a \rho_{ij}$ en plus de l'information portée par c_j. Ce qui signifie que la description d'un concept comporte toute l'information nécessaire à la description de son père plus une certaine quantité d'information qui lui est propre.

Ainsi, il apparaît que le calcul du contenu informationnel global d'un ensemble de concepts $\mathcal{C}_i$ correspond à la somme des quantités d'information propres à la description de chaque subsumant non strict d'au moins un des concepts de $\mathcal{C}_i$. Cette quantité d'information propre à un concept est obtenue en faisant la différence entre son contenu informationnel et celui de son père (cf. figure 6.2 pour un exemple) :

$$\begin{aligned} \psi^{\cup}(\emptyset) &= 0 \\ \psi^{\cup}(\mathcal{C}_i) &= \psi(c_0) + \sum_{c_x \in \cup^{\sqsubset}_{\mathcal{C}_i} - \{c_0\}} \psi(c_x) - \psi(c_x^{\star}) \\ &= \psi(c_0) + \sum_{c_x \in \cup^{\sqsubset}_{\mathcal{C}_i} - \{c_0\}} -\log \frac{\mathrm{P}(c_x)}{\mathrm{P}(c_x^{\star})} \end{aligned} \tag{6.1}$$

Cas d'une hiérarchie

Dans le cas d'une hiérarchie dans laquelle un concept peut avoir plusieurs parents, l'équation 6.1 ne s'applique plus. Le raisonnement reste néanmoins le même : il s'agit toujours de faire la somme des quantités d'information propres à la description de chaque subsumant non strict d'au moins un des concepts de $\mathcal{C}_i$. En revanche, un concept c_i pouvant avoir plusieurs parents, il s'agit cette fois-ci de faire la différence entre le contenu informationnel de c_i et le contenu



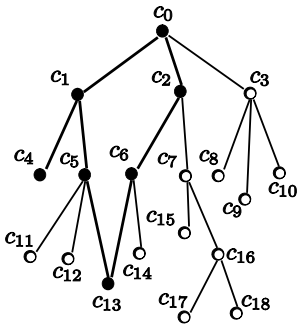
$$\mathcal{C}_i = \{c_{15}, c_{18}\}$$

$$\begin{aligned}\psi^{\cup}(\mathcal{C}_i) &= \psi(c_{18}) - \psi(c_{16}) + \psi(c_{16}) - \psi(c_7) \\ &\quad + \psi(c_{15}) - \psi(c_7) + \psi(c_7) - \psi(c_2) \\ &\quad + \psi(c_2) - \psi(c_0) + \psi(c_0) \\ &= \psi(c_{18}) + \psi(c_{15}) - \psi(c_7)\end{aligned}$$

FIG. 6.2 – Contenu informationnel global d'un sous-ensemble de concepts d'un arbre de subsomption

informationnel global de ses parents (cf. figure 6.3 pour un exemple) :

$$\begin{aligned}\psi^{\cup}(\emptyset) &= 0 \\ \psi^{\cup}(\mathcal{C}_i) &= \sum_{c_x \in \cup \overline{\mathcal{C}_i}^{\sqsubset}} \psi(c_x) - \psi^{\cup}\left(c_x^{\prec}\right) \\ &= \psi(c_0) + \sum_{c_x \in \cup \overline{\mathcal{C}_i}^{\sqsubset !}} -\log \frac{\mathrm{P}(c_x)}{\mathrm{P}(c_x^{\star})} + \sum_{c_x \in \cup \overline{\mathcal{C}_i}^{\sqsubset \natural}} \psi(c_x) - \psi^{\cup}\left(c_x^{\prec}\right)\end{aligned} \tag{6.2}$$



$$\mathcal{C}_i = \{c_4, c_{13}\}$$

$$\begin{aligned}\psi^{\cup}(\mathcal{C}_i) &= \psi(c_{13}) - \psi^{\cup}(\{c_5, c_6\}) \\ &\quad + \psi(c_4) - \psi(c_1) + \psi(c_6) - \psi(c_2) \\ &\quad + \psi(c_5) - \psi(c_1) + \psi(c_1) - \psi(c_0) \\ &\quad + \psi(c_2) - \psi(c_0) + \psi(c_0)\end{aligned}$$

$$\begin{aligned}\psi^{\cup}(\{c_5, c_6\}) &= \psi(c_5) - \psi(c_1) + \psi(c_1) - \psi(c_0) \\ &\quad + \psi(c_6) - \psi(c_2) + \psi(c_2) - \psi(c_0) \\ &\quad + \psi(c_0)\end{aligned}$$

$$\psi^{\cup}(\mathcal{C}_i) = \psi(c_{13}) + \psi(c_4) - \psi(c_1)$$

FIG. 6.3 – Contenu informationnel global d'un ensemble de concepts d'une hiérarchie de subsomption avec héritage multiple

Le calcul du contenu informationnel global d'un ensemble de concepts nécessite donc une application récursive de l'équation 6.2 comme on peut le voir sur l'exemple avec le calcul de $\psi^{\cup}(\{c_5, c_6\})$.

Remarque. L'information apportée par un concept englobe celle de tous ses subsumants. Le contenu informationnel $\psi^{\cup}(\mathcal{C}_i)$ d'un ensemble $\mathcal{C}_i$ de concepts est donc égal au contenu informationnel du sous-ensemble $\Gamma(\mathcal{C}_i)$ des concepts n'ayant pas de subsumés parmi les concepts de $\mathcal{C}_i$:

$$\begin{aligned} &\psi^{\cup}(\mathcal{C}_i) = \psi^{\cup}(\Gamma(\mathcal{C}_i)) \\ &\text{avec, } \Gamma(\mathcal{C}_i) = \mathcal{C}_i - \cup_{\mathcal{C}_i}^{\sqsubset} \end{aligned} \tag{6.3}$$

6.3.3 Contenu informationnel partagé

Cette approche nous permet également d'évaluer la quantité d'information partagée par les descriptions respectives de deux concepts. Nous notons $\psi^{\cap}(\mathcal{C}_i)$ la quantité d'information partagée par un ensemble de concepts $\mathcal{C}_i$.

Cas d'un arbre

Si on prend l'exemple de la figure 6.4, c_{15} et c_{18} ont trois subsumants communs (c_0,c_2 et c_7). Dans cet exemple, c_0 est une racine virtuelle ce qui explique que son contenu informationnel ne soit pas représenté puisqu'il est nul. L'information nécessaire à la description de c_7 englobe celle nécessaire à la description de c_2. Selon cette hiérarchie, le contenu informationnel de c_7 constitue donc le contenu informationnel partagé par c_{15} et c_{18}.

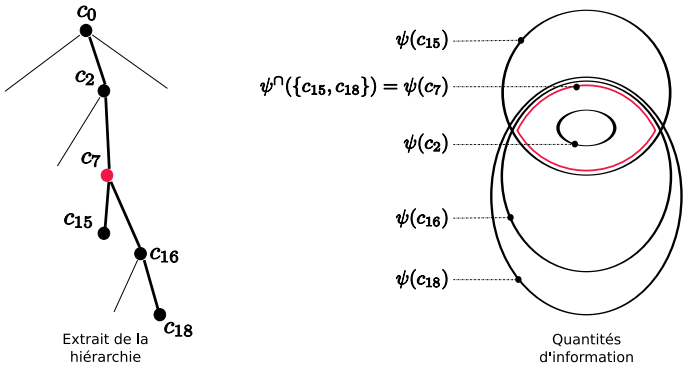


FIG. 6.4 – Subsumant commun le plus spécifique et quantité d'information partagée

En résumé, dans le cas d'un arbre, le contenu informationnel partagé $\psi^{\cap}(\{c_i, c_j\})$ par les concepts c_i et c_j correspond au contenu informationnel de leur subsumant commun le plus spécifique $c_{ij} \in \mathcal{C}$ (cf. figure 6.5) :

$$\begin{aligned} &\psi^{\cap}(\{c_i, c_j\}) = \psi(c_{ij}) \\ &\text{avec, } \psi(c_{ij}) = \max_{c_x \in c_i^{\sqsubseteq} \cap c_j^{\sqsubseteq}} \psi(c_x) \end{aligned} \tag{6.4}$$

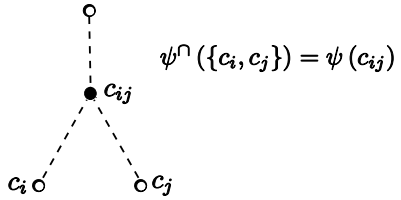


FIG. 6.5 – Contenu informationnel partagé dans un arbre

De manière plus générale, le contenu informationnel partagé $\psi^{\cap}(\mathcal{C}_i)$ par un sous-ensemble de concepts $\mathcal{C}_i \subseteq \mathcal{C}$ de cardinalité quelconque correspond à la quantité d'information de leur subsumant commun le plus spécifique :

$$\psi^{\cap}(\mathcal{C}_i) = \max_{c_x \in \cap \overline{\mathcal{C}_i}^{\sqsubseteq}} \psi(c_x) \tag{6.5}$$

Cas d'une hiérarchie

Dans le cas d'un arbre, l'information apportée par c_{ij} englobe l'information apportée par tous les autres subsumants communs. En revanche, pour une hiérarchie, cette constatation ne tient plus, puisque deux subsumants communs peuvent ne pas être subsumants l'un de l'autre (cf. figure 6.6).

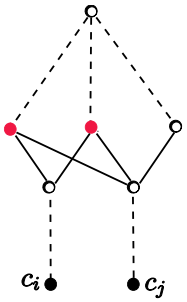


FIG. 6.6 – Cas où deux subsumants communs ne sont pas subsumants l'un de l'autre

De manière plus générale, la quantité d'information partagée $\psi^{\cap}(\{c_i, c_j\})$ par deux concepts c_i et c_j correspond au contenu informationnel $\psi^{\cup}\left(c_i^{\overline{\sqsubseteq}} \cap c_j^{\overline{\sqsubseteq}}\right)$ de l'ensemble des subsumants communs à c_i et c_j comme le montre la figure 6.7.

Pour le calcul du contenu informationnel partagé $\psi^{\cap}(\{c_i, c_j\})$ par c_i et c_j, on peut se restreindre à considérer l'ensemble des subsumants communs n'étant pas subsumés par un autre subsumant commun (cf. figure 6.8) :

$$\psi^{\cap}(\{c_i, c_j\}) = \psi^{\cup}\left(\Gamma\left(c_i^{\overline{\sqsubseteq}} \cap c_j^{\overline{\sqsubseteq}}\right)\right) \tag{6.6}$$

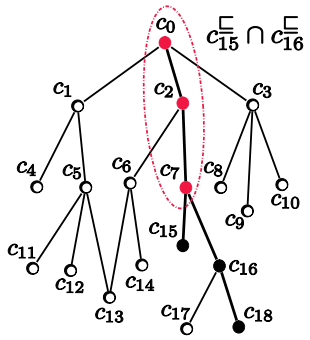


FIG. 6.7 – Exemple d'un ensemble de subsumants communs les plus spécifiques

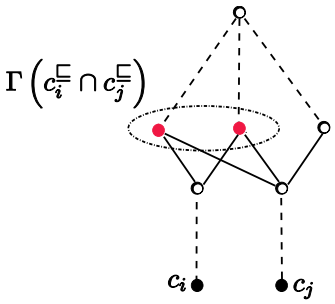


FIG. 6.8 – Contenu informationnel partagé dans une hiérarchie avec héritage multiple

Ce résultat se généralise à la quantité d'information partagée $\psi^{\cap}(\mathcal{C}_i)$ par un ensemble de concepts $\mathcal{C}_i$:

$$\psi^{\cap}(\mathcal{C}_i) = \psi^{\cup}\left(\Gamma\left(\cap_{\mathcal{C}_i}^{\sqsubseteq}\right)\right) \tag{6.7}$$

6.3.4 Reformulation du contenu informationnel global

Nous venons de définir le contenu informationnel global d'un ensemble de concepts, ainsi que le contenu informationnel partagé par un ensemblde de concepts. Nous menons maintenant un raisonnement complémentaire pour simplifier l'implémentation du contenu informationnel global $\psi^{\cup}$ dans une hiérarchie.

Le contenu informationnel global $\psi^{\cup}$ d'un ensemble de concepts ne peut pas se résumer à la somme des contenus informationnels de chaque concept de cet ensemble du fait que certains concepts peuvent partager de l'information. Et, comme on le voit sur les figures 6.2 et 6.3, il est possible d'exprimer de manière beaucoup plus concise le contenu informationnel global. Nous cherchons donc à

exprimer le contenu informationnel global $\psi^{\cup}(\mathcal{C}_i)$ d'un ensemble de concepts $\mathcal{C}_i$ en fonction d'un minimum de contenus informationnels de concepts de $\mathcal{C}$.

Nous considérons maintenant successivement le calcul du contenu informationnel global d'un ensemble vide, d'un singleton $\{c_i\}$, d'une paire $\{c_i, c_j\}$ et enfin celui d'un ensemble de trois concepts $\{c_i, c_j, c_k\}$. Cela nous permettra de généraliser le calcul à tout ensemble de concepts $\mathcal{C}_i$ quelque soit son cardinal. Trivialement, le contenu informationnel global d'un ensemble vide est nul et celui d'un singleton $\{c_i\}$ est égal au contenu informationnel du concept c_i :

$$\psi^{\cup}(\emptyset) = 0$$
$$\psi^{\cup}(\{c_i\}) = \psi(c_i)$$

Pour rendre compte du contenu informationnel de la paire $\{c_i, c_j\}$, il est assez intuitif (cf. figure 6.9) d'ajouter le contenu informationnel de chaque concept et de soustraire le contenu informationnel partagé :

$$\psi^{\cup}(\{c_i, c_j\}) = \psi(c_i) + \psi(c_j) - \psi^{\cap}(\{c_i, c_j\}) \tag{6.8}$$

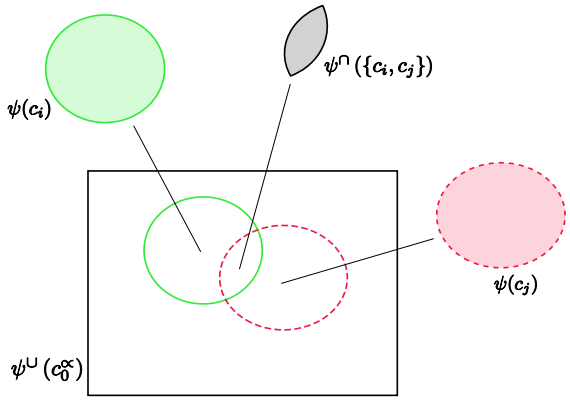


FIG. 6.9 – Représentation des quantités d'information pour une paire de concepts $\{c_i, c_j\}$

Nous décomposons en deux étapes ce que nous venons naturellement de faire en prenant en compte à la n$^{\text{ème}}$ étape uniquement le contenu informationnel des n-combinaisons[1] de la paire de concepts considérée :

1. Nous ajoutons la quantité d'information $\psi(c_i)$ et $\psi(c_j)$ de chaque concept.
2. Nous soustrayons la quantité d'information partagée $\psi^{\cap}(\{c_i, c_j\})$:
   (a) Nous devons comptabiliser une fois cette quantité d'information.
   (b) A la première étape, nous avons indirectement ajouté deux fois (nombre de 1-combinaisons de l'ensemble à deux éléments $\{c_i, c_j\}$, c'est-à-dire $\binom{2}{1} = 2$) cette quantité d'information.

[1] Une n-combinaison d'un ensemble E est une partie de E contenant n éléments

On peut alors réécrire l'équation précédente de manière détaillée comme suit :

$$\begin{aligned}\psi^{\cup}(\{c_i,c_j\}) = \quad & 1\cdot[\psi(c_i)+\psi(c_j)] \\ & +\left[1-1\cdot\binom{2}{1}\right]\cdot\psi^{\cap}(\{c_i,c_j\})\end{aligned} \tag{6.9}$$

Le calcul de la quantité d'information d'un ensemble de trois concepts $\{c_i, c_j, c_k\}$ est déjà moins une évidence. Les quantités d'information mises en jeu sont $\psi(c_i)$, $\psi(c_j)$, $\psi(c_k)$, $\psi^{\cap}(\{c_i,c_j\})$, $\psi^{\cap}(\{c_i,c_k\})$, $\psi^{\cap}(\{c_j,c_k\})$ et $\psi^{\cap}(\{c_i,c_j,c_k\})$ (cf. figure 6.10). Nous faisons une décomposition en trois étapes en prenant en compte à la n$^{\text{ème}}$ étape uniquement le contenu informationnel des n-combinaisons de l'ensemble de trois concepts considéré :

1. Nous ajoutons la quantité d'information $\psi(c_i)$, $\psi(c_j)$ et $\psi(c_k)$ de chaque concept
2. Nous soustrayons les quantités d'information $\psi^{\cap}(\{c_i,c_j\})$, $\psi^{\cap}(\{c_i,c_k\})$ et $\psi^{\cap}(\{c_j,c_k\})$:
   (a) Nous devons comptabiliser une fois ces quantités d'information.
   (b) A l'étape précédente, nous avons indirectement ajouté $\binom{2}{1}=2$ fois ces quantités d'information
3. Nous ajoutons la quantité d'information $\psi^{\cap}(\{c_i,c_j,c_k\})$:
   (a) Nous devons comptabiliser une fois cette quantité d'information.
   (b) A la première étape, nous avons indirectement ajouté $\binom{3}{1}=3$ fois cette quantité d'information.
   (c) A la seconde étape, nous avons indirectement soustrait $\binom{3}{2}=3$ fois cette quantité d'information.

En résumé, le contenu informationnel global pour un ensemble de trois concepts se calcul comme suit :

$$\begin{aligned}&\psi^{\cup}(\{c_i,c_j,c_k\}) \\ &\quad = 1\cdot[\psi(c_i)+\psi(c_j)] \\ &\qquad +\left[1-1\cdot\binom{2}{1}\right]\cdot[\psi^{\cap}(\{c_i,c_j\})+\psi^{\cap}(\{c_i,c_k\})+\psi^{\cap}(\{c_j,c_k\})] \\ &\qquad +\left[1-1\cdot\binom{3}{1}-(-1)\cdot\binom{3}{2}\right]\cdot\psi^{\cap}(\{c_i,c_j,c_k\}) \\ &\quad = \psi(c_i)+\psi(c_j)+\psi(c_k)-\psi^{\cap}(\{c_i,c_j\})-\psi^{\cap}(\{c_i,c_k\}) \\ &\qquad -\psi^{\cap}(\{c_j,c_k\})+\psi^{\cap}(\{c_i,c_j,c_k\})\end{aligned}$$

Nous généralisons cette équation pour un ensemble de concepts $\mathcal{C}_i$ ($\mathcal{P}_k(E)$ ensemble des parties à k éléments de l'ensemble E) :

$$\begin{aligned}\psi^{\cup}(\mathcal{C}_i) \quad & = \sum_{k=1}^{|\mathcal{C}_i|}\left(u_k\cdot\sum_{\mathcal{C}_x\in\mathcal{P}_k(\mathcal{C}_i)}\psi^{\cap}(\mathcal{C}_x)\right) \\ & \text{avec,} \left\{\begin{array}{ll} u_1 & =1 \\ u_n & =1-\sum_{i=1}^{n-1}u_i\cdot\binom{n}{i}\end{array}\right\}_{n>1}\end{aligned} \tag{6.10}$$

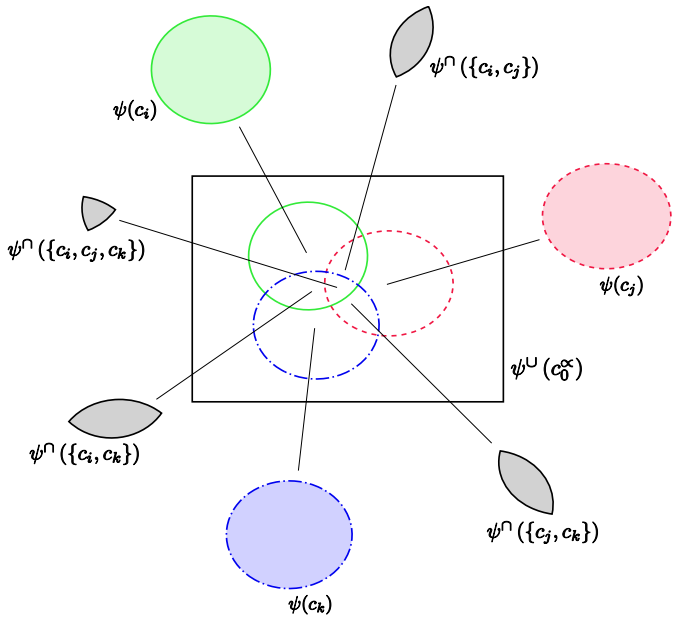


FIG. 6.10 – Représentation des quantités d'information pour un triplet de concepts $\{c_i, c_j, c_k\}$

Les coefficients u_n nécessaires sont liés par une relation de récurrence linéaire complète[2] tout à fait particulière. En regardant les premiers termes $u_1 = 1$, $u_2 = -1$, $u_3 = 1$, $u_4 = -1$, $u_5 = 1$, etc. de la suite u_n, on s'aperçoit que nous avons à faire à une suite alternée. D'où la proposition suivante qui permet une implémentation efficace du contenu informationnel global :

$$\psi^{\cup}(\mathcal{C}_i) = \sum_{k=1}^{|\mathcal{C}_i|} \left((-1)^{k+1} \cdot \sum_{\mathcal{C}_x \in \mathcal{P}_k(\mathcal{C}_i)} \psi^{\cap}(\mathcal{C}_x) \right) \tag{6.11}$$

On vérifie aisément que u_n s'écrit $u_n = (-1)^{n+1}$, $n \geq 1$.

Démonstration 6.1 *On définit les suites u_n et v_n comme suit :*

$$u_n = \begin{cases} 1 & , n = 1 \\ 1 - \sum_{i=1}^{n-1} u_i \cdot \binom{n}{i} & , n > 1 \end{cases}$$

$$v_n = (-1)^{n+1}$$

On cherche à montrer par récurrence l'égalité $u_n = v_n$.

[2]Il s'agit d'une relation de récurrence linéaire complète (ou forte) du fait que le terme de rang n est une combinaison linéaire de l'ensemble des termes de rang 1 à $n-1$

Au rang 1, $u_n = v_n$ *est vérifiée :*

$$\begin{aligned} v_1 &= (-1)^{1+1} \\ &= u_1 \end{aligned}$$

On suppose que $u_n = v_n$ *est vérifiée aux rangs* 1 *à* $n-1$ *et on cherche à montrer que dans ce cas,* $u_n = v_n$ *est vrai au rang* n *:*

$$\begin{aligned} u_n &= 1 - \sum_{i=1}^{n-1} u_i \cdot \binom{n}{i} \\ &= 1 - \sum_{i=1}^{n-1} (-1)^{i+1} \cdot \binom{n}{i} \\ &= 1 - \left(\left[\sum_{i=0}^{n} (-1)^{i+1} \cdot \binom{n}{i} \right] - \left[(-1)^1 \cdot \binom{n}{0} \right] - \left[(-1)^{n+1} \cdot \binom{n}{n} \right] \right) \\ &= (-1)^{n+1} + \sum_{i=0}^{n} (-1)^i \cdot \binom{n}{i} \end{aligned}$$

Or, une des propriétés du triangle de Pascal est que la somme alternée des termes d'une ligne est nulle :

$$\sum_{i=0}^{n} (-1)^i \cdot \binom{n}{i} = 0$$

D'où $u_n = v_n$ *est vérifiée au rang* n*.*

6.3.5 Choix de l'unité d'information

Dans le cas d'une hiérarchie avec héritage multiple, on ne peut pas appréhender la spécificité de tous les concepts par leur profondeur puisqu'il existe parfois plusieurs chemins de taille différente pour rejoindre la racine. Il est cependant intéressant que tous ceux pour lesquels il existe un seul chemin aient une spécificité de l'ordre de leur profondeur. Il faut donc toujours utiliser l'approximation $\widehat{\mathrm{P}}_p$ avec $a = \kappa$ sauf que l'on considère la quantité d'information de l'ensemble des concepts $\mathcal{C}' = \{c | c \in \mathcal{C} \wedge c^{\sqsupseteq} - \{c_0\} = c^{\sqsupseteq !}\}$ pour lesquels il n'y a qu'un chemin vers la racine :

$$\begin{aligned} \widehat{\mathrm{P}}_p(\mathcal{C}') &= \psi(c_0) + \sum_{c_x \in \mathcal{C}' - \{c_0\}} -\log_\kappa \frac{1}{\kappa} \\ &= \psi(c_0) + |\mathcal{C}' - \{c_0\}| \\ &= \psi(c_0) + |\mathcal{C}'| - 1 \end{aligned}$$

On retrouve ainsi la base du logarithme qui est la moyenne géométrique des inverses des probabilités conditionnelles $P(c_x/c_x^\star)$ restreinte aux concepts $c_x \in \mathcal{C}'$ pour lesquels il n'existe qu'un chemin vers la racine :

$$a = \sqrt[|\mathcal{C}' - \{c_0\}|]{\prod_{c_x \in \mathcal{C}' - \{c_0\}} \frac{\mathrm{P}(c_x^\star)}{\mathrm{P}(c_x)}}$$

6.4 Adaptation des approximations

L'approche ascendante dont relèvent les approximations $\widehat{\mathrm{P}}_h$ et $\widehat{\mathrm{P}}_g$ n'est pas affectée par le relâchement de la contrainte qui fait d'un arbre une hiérarchie. En revanche, la notion de profondeur n'a plus de sens puisqu'il peut y avoir plusieurs chemins pour rejoindre la racine ce qui nous oblige à étendre l'approximation $\widehat{\mathrm{P}}_p$. Dans le cas de l'approximation $\widehat{\mathrm{P}}_s$, l'hypothèse d'une disjonction systématique doit être remise en cause le cas échéant.

6.4.1 Approximation $\widehat{\mathrm{P}}_p$

Du fait de la possible multiplicité des chemins entre un concept et la racine, la notion de profondeur n'est plus appropriée. Pour généraliser l'approximation $\widehat{\mathrm{P}}_p$ à une hiérarchie, nous reprenons l'égalité entre le contenu informationnel d'un concept et la somme des quantités d'information apportées par chaque subsumant non strict de ce concept. Chaque concept c_i n'ayant qu'un père apporte une information vis-à-vis de celui-ci égale à $\psi_p(c_i) - \psi_p(c_i^\star) = -\log_a \frac{1}{\kappa}$. Nous envisageons deux alternatives pour évaluer la quantité d'information $-\log_a \rho$ apportée par un concept ayant plusieurs parents : nous pouvons considérer

1. qu'un fils ayant plusieurs parents est le fruit de la réunion de ses parents et n'apporte aucune information propre en prenant $\rho = 1$;
2. qu'il apporte une certaine quantité d'information supplémentaire (hypothèse parfois plus réaliste dans le contexte des ontologies) en prenant $0 < \rho < 1$.

Cette quantité d'information supplémentaire pourraît être équivalente à la quantité d'information ajoutée lors d'une spécialisation classique en prenant $\rho = \frac{1}{\kappa}$. Cela nous permet de redéfinir l'approximation $\widehat{\mathrm{P}}_p$ (cf. figure 6.11 pour l'exemple) :

$$
\begin{aligned}
\psi_p(c_i) &= \sum_{c_x \in c_i^{\sqsubseteq}} \psi_p(c_x) - \psi_p^{\cup}\left(c_x^{\prec}\right) \\
-\log_a \widehat{\mathrm{P}}_p(c_i) &= -\log_a \widehat{\mathrm{P}}(c_0) + \sum_{c_x \in c_i^{\sqsubseteq !}} -\log_a \frac{1}{\kappa} + \sum_{c_x \in c_i^{\sqsubseteq \natural}} -\log_a \rho \\
-\log_a \widehat{\mathrm{P}}_p(c_i) &= -\log_a \left(\widehat{\mathrm{P}}(c_0) \cdot \prod_{c_x \in c_i^{\sqsubseteq !}} \frac{1}{\kappa} \cdot \prod_{c_x \in c_i^{\sqsubseteq \natural}} \rho \right) \\
\widehat{\mathrm{P}}_p(c_i) &= \widehat{\mathrm{P}}(c_0) \cdot \left(\frac{1}{\kappa}\right)^{\left|c_i^{\sqsubseteq !}\right|} \cdot \rho^{\left|c_i^{\sqsubseteq \natural}\right|}
\end{aligned}
\tag{6.12}
$$

Comme dans la définition restreinte à un arbre, nous pouvons fixer κ comme unité d'information pour le contenu informationnel issue de l'approximation $\widehat{\mathrm{P}}_p$. Suivant le choix du paramètre ρ, le contenu informationnel s'exprime comme

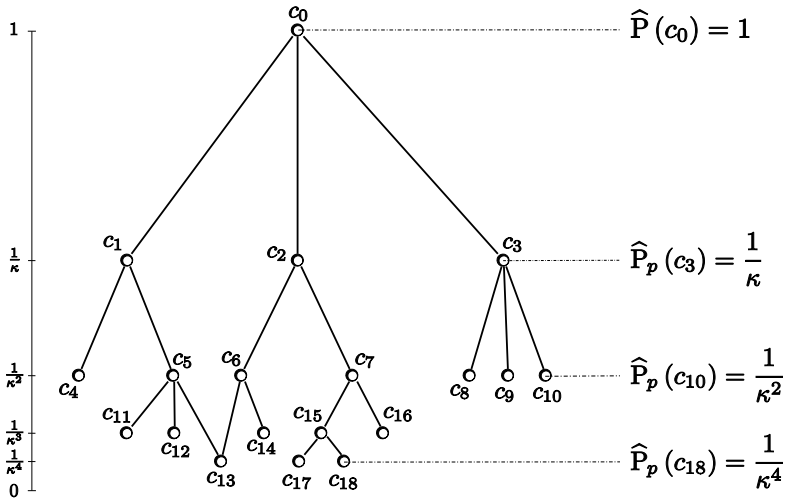


FIG. 6.11 – Application de l'approximation $\widehat{\mathrm{P}}_p$ avec $\widehat{\mathrm{P}}(c_0) = 1$ et $\rho = 1$

suit :

$$\psi_p(c_i) = \left|c_i^{\sqsubset!}\right| + \psi(c_0) \qquad \text{, pour } \rho = 1$$

$$\psi_p(c_i) = \left|c_i^{\sqsubset}\right| + \psi(c_0) \qquad \text{, pour } \rho = \frac{1}{\kappa}$$

Remarque. Pour généraliser la profondeur, on s'aperçoit ici qu'il ne s'agit pas de prendre le plus court ou le plus long chemin, mais qu'il faut considérer les contributions de tous les subsumants.

6.4.2 Approximation $\widehat{\mathrm{P}}_s$

Pour généraliser cette approximation à une hiérarchie, il est nécessaire d'adapter la contrainte d'uniformité des frères que doit respecter cette approximation :

$$\forall c_i, c_j \in \mathcal{C}, c_i^{\prec} = c_j^{\prec} \Longrightarrow |\mathcal{E}_i| = |\mathcal{E}_j| \qquad \text{(uniformité des frères)}$$

Cette approximation repose sur l'hypothèse d'une disjonction de l'extension des concepts frères. Lors de la généralisation de cette approximation, nous sommes confrontés à la remise en cause de cette hypothèse lorsque certains frères ont des subsumés en commun. Ces subsumés en commun sont des concepts ayant plusieurs parents et dont la probabilité associée doit être exprimée en fonction de leurs parents. Un fils c_i ayant plusieurs parents est considéré comme le fruit de la réunion de ses parents et apporte une certaine quantité d'information propre $-\log_a \rho$ ($\rho \in [0;1]$) avec ρ la probabilité pour une instance d'appartenir au concept c_i sachant qu'elle appartient à ses parents. Cette probabilité peut être maximale ($\rho = 1$) ou bien fixée à une autre valeur de manière arbitraire.

Nous pouvons également la calculer en faisant la moyenne arithmétique des apports d'information des concepts de la hiérarchie n'ayant qu'un seul père, ce qui amène à calculer la moyenne géométrique de la probabilité qu'une instance appartienne à un concept sachant qu'elle appartient à son père :

$$
\begin{aligned}
-\log_a \rho &= \frac{\sum_{c_x \in c_0^{\sqsupset !}} \psi\left(c_x\right) - \psi\left(c_x^\star\right)}{\left|c_0^{\sqsupset !}\right|} \\
-\log_a \rho &= \frac{\sum_{c_x \in c_0^{\sqsupset !}} -\log_a \mathrm{P}\left(c_x\right) + \log_a \mathrm{P}\left(c_x^\star\right)}{\left|c_0^{\sqsupset !}\right|} \\
-\log_a \rho &= \frac{\sum_{c_x \in c_0^{\sqsupset !}} -\log_a \frac{\mathrm{P}\left(c_x\right)}{\mathrm{P}\left(c_x^\star\right)}}{\left|c_0^{\sqsupset !}\right|} \\
-\log_a \rho &= \frac{-\log_a \prod_{c_x \in c_0^{\sqsupset !}} \frac{\mathrm{P}\left(c_x\right)}{\mathrm{P}\left(c_x^\star\right)}}{\left|c_0^{\sqsupset !}\right|} \\
\rho &= \left(\prod_{c_x \in c_0^{\sqsupset !}} \frac{\mathrm{P}\left(c_x\right)}{\mathrm{P}\left(c_x^\star\right)}\right)^{\frac{1}{\left|c_0^{\sqsupset !}\right|}} \\
\rho &= \sqrt[\left|c_0^{\sqsupset !}\right|]{\prod_{c_x \in c_0^{\sqsupset !}} \frac{\mathrm{P}\left(c_x\right)}{\mathrm{P}\left(c_x^\star\right)}}
\end{aligned}
$$

Quelque soit la valeur de ρ que l'on fixe, on évalue le contenu informationnel d'un concept ayant plusieurs parents comme suit :

$$\psi_s\left(c_i\right) = \psi_s^{\cup}\left(c_i^{\prec}\right) - \log_a \rho$$

Ajoutées aux contraintes induites par la propriété de complétude et par l'équirépartition des instances d'un père vers ses fils, ces contraintes sur les fils non exclusifs[3] nous permettent de générer un système d'équations. La définition de ce système d'équations passe par la prise en compte de trois types de contraintes :

1. La probabilité d'un concept noeud c_i est égale à la somme des probabilités associées aux feuilles qu'il susbsume. (on engendre autant d'équations que de concepts noeuds) :

   $$\widehat{\mathrm{P}}\left(c_i\right) = \sum_{c_j \in c_i^{\propto}} \widehat{\mathrm{P}}\left(c_j\right)$$

   L'équation obtenue à partir de la racine est essentielle, puisqu'elle garantie l'unicité de la solution (du fait de la connaissance de $\widehat{\mathrm{P}}\left(c_0\right)$) telle que les probabilités des concepts soient dans l'intervalle $[0;1]$. Le relâchement de la complétude pourra être mis en oeuvre en ajoutant effectivement ϵ concepts fils à chaque concept noeud. Tous les concepts ajoutés sont autant de concepts feuille supplémentaires.

[3] Un fils exclusif est un concept fils n'ayant pas d'autre père.

Sur notre exemple, en prenant $\widehat{\mathrm{P}}(c_0) = 1$ et $\epsilon = 0$, on obtient les huit équations suivantes :

$$\begin{array}{l}
\widehat{\mathrm{P}}_s(c_4) + \widehat{\mathrm{P}}_s(c_{11}) + \widehat{\mathrm{P}}_s(c_{12}) + \widehat{\mathrm{P}}_s(c_{13}) + \widehat{\mathrm{P}}_s(c_{14}) + \widehat{\mathrm{P}}_s(c_{17}) \\
+\widehat{\mathrm{P}}_s(c_{18}) + \widehat{\mathrm{P}}_s(c_{16}) + \widehat{\mathrm{P}}_s(c_8) + \widehat{\mathrm{P}}_s(c_9) + \widehat{\mathrm{P}}_s(c_{10}) = 1 \\
\widehat{\mathrm{P}}_s(c_{11}) + \widehat{\mathrm{P}}_s(c_{12}) + \widehat{\mathrm{P}}_s(c_{13}) = \widehat{\mathrm{P}}_s(c_5) \\
\widehat{\mathrm{P}}_s(c_{13}) + \widehat{\mathrm{P}}_s(c_{14}) = \widehat{\mathrm{P}}_s(c_6) \\
\widehat{\mathrm{P}}_s(c_4) + \widehat{\mathrm{P}}_s(c_{11}) + \widehat{\mathrm{P}}_s(c_{12}) + \widehat{\mathrm{P}}_s(c_{13}) = \widehat{\mathrm{P}}_s(c_1) \\
\widehat{\mathrm{P}}_s(c_{13}) + \widehat{\mathrm{P}}_s(c_{14}) + \widehat{\mathrm{P}}_s(c_{17}) + \widehat{\mathrm{P}}_s(c_{18}) + \widehat{\mathrm{P}}_s(c_{16}) = \widehat{\mathrm{P}}_s(c_2) \\
\widehat{\mathrm{P}}_s(c_8) + \widehat{\mathrm{P}}_s(c_9) + \widehat{\mathrm{P}}_s(c_{10}) = \widehat{\mathrm{P}}_s(c_3) \\
\widehat{\mathrm{P}}_s(c_{17}) + \widehat{\mathrm{P}}_s(c_{18}) + \widehat{\mathrm{P}}_s(c_{16}) = \widehat{\mathrm{P}}_s(c_7) \\
\widehat{\mathrm{P}}_s(c_{17}) + \widehat{\mathrm{P}}_s(c_{18}) = \widehat{\mathrm{P}}_s(c_{15})
\end{array}$$

2. Les probabilités associées à un ensemble de n ($n > 1$) fils exclusifs $\{c_{i,1}, c_{i,2}, c_{i,3}, \ldots, c_{i,n-1}, c_{i,n}\}$ d'un même père sont égales. Donc chaque ensemble de n ($n > 1$) fils exclusifs d'un même père, engendre $n - 1$ équations :

$$\begin{array}{l}
\widehat{\mathrm{P}}_s(c_{i,1}) = \widehat{\mathrm{P}}_s(c_{i,2}) \\
\widehat{\mathrm{P}}_s(c_{i,2}) = \widehat{\mathrm{P}}_s(c_{i,3}) \\
\ldots \\
\widehat{\mathrm{P}}_s(c_{i,n-1}) = \widehat{\mathrm{P}}_s(c_{i,n})
\end{array}$$

Sur notre exemple, on obtient les neuf équations suivantes :

$$\begin{array}{l}
\widehat{\mathrm{P}}_s(c_1) = \widehat{\mathrm{P}}_s(c_2) \\
\widehat{\mathrm{P}}_s(c_2) = \widehat{\mathrm{P}}_s(c_3) \\
\widehat{\mathrm{P}}_s(c_4) = \widehat{\mathrm{P}}_s(c_5) \\
\widehat{\mathrm{P}}_s(c_6) = \widehat{\mathrm{P}}_s(c_7) \\
\widehat{\mathrm{P}}_s(c_{15}) = \widehat{\mathrm{P}}_s(c_{16}) \\
\widehat{\mathrm{P}}_s(c_{11}) = \widehat{\mathrm{P}}_s(c_{12}) \\
\widehat{\mathrm{P}}_s(c_{17}) = \widehat{\mathrm{P}}_s(c_{18}) \\
\widehat{\mathrm{P}}_s(c_8) = \widehat{\mathrm{P}}_s(c_9) \\
\widehat{\mathrm{P}}_s(c_9) = \widehat{\mathrm{P}}_s(c_{10})
\end{array}$$

3. Une contrainte associée à chaque fils c_i non exclusif donne lieu à autant d'équations non linéaires :

$$\psi_s(c_i) = \psi_s^{\cup}\left(c_i^{\prec}\right) - \log_a \rho$$

Sur notre exemple, en prenant $\rho = 1$ et $\widehat{\mathrm{P}}(c_0) = 1$, on obtient l'équation suivante ($\psi_s(c_{13}) = \psi_s^{\cup}(\{c_5, c_6\}) = \widehat{\mathrm{P}}_s(c_5) + \widehat{\mathrm{P}}_s(c_6) - \widehat{\mathrm{P}}(c_0)$) :

$$\widehat{\mathrm{P}}_s(c_{13}) = \widehat{\mathrm{P}}_s(c_5) \cdot \widehat{\mathrm{P}}_s(c_6)$$

La complexité du système obtenu est grande dans la mesure où le nombre de variables et d'équations par la même occasion est égale au nombre de concepts de la hiérarchie privée de la racine ($|\mathcal{C} - \{c_0\}| = 18$ dans notre exemple). Après

développement, le système à résoudre est le suivant :

$$\begin{cases}
\widehat{\mathrm{P}}_s(c_{13})^2 - 34 \cdot \widehat{\mathrm{P}}_s(c_{13}) + 1 = 0 \\
\widehat{\mathrm{P}}_s(c_{14}) = \dfrac{1 - 5 \cdot \widehat{\mathrm{P}}_s(c_{13})}{6} \\
\widehat{\mathrm{P}}_s(c_{11}) = \dfrac{\widehat{\mathrm{P}}_s(c_{14})}{2} \\
\widehat{\mathrm{P}}_s(c_4) = \widehat{\mathrm{P}}_s(c_{14}) + \widehat{\mathrm{P}}_s(c_{13}) \\
\widehat{\mathrm{P}}_s(c_{16}) = \dfrac{\widehat{\mathrm{P}}_s(c_4)}{2} \\
\widehat{\mathrm{P}}_s(c_{17}) = \dfrac{\widehat{\mathrm{P}}_s(c_{16})}{2} \\
\widehat{\mathrm{P}}_s(c_8) = \dfrac{2 \cdot \widehat{\mathrm{P}}_s(c_4)}{3} \\
\widehat{\mathrm{P}}_s(c_{11}) = \widehat{\mathrm{P}}_s(c_{12}) \\
\widehat{\mathrm{P}}_s(c_{17}) = \widehat{\mathrm{P}}_s(c_{18}) \\
\widehat{\mathrm{P}}_s(c_8) = \widehat{\mathrm{P}}_s(c_9) \\
\widehat{\mathrm{P}}_s(c_9) = \widehat{\mathrm{P}}_s(c_{10}) \\
\widehat{\mathrm{P}}_s(c_{11}) + \widehat{\mathrm{P}}_s(c_{12}) + \widehat{\mathrm{P}}_s(c_{13}) = \widehat{\mathrm{P}}_s(c_5) \\
\widehat{\mathrm{P}}_s(c_{13}) + \widehat{\mathrm{P}}_s(c_{14}) = \widehat{\mathrm{P}}_s(c_6) \\
\widehat{\mathrm{P}}_s(c_4) + \widehat{\mathrm{P}}_s(c_{11}) + \widehat{\mathrm{P}}_s(c_{12}) + \widehat{\mathrm{P}}_s(c_{13}) = \widehat{\mathrm{P}}_s(c_1) \\
\widehat{\mathrm{P}}_s(c_{13}) + \widehat{\mathrm{P}}_s(c_{14}) + \widehat{\mathrm{P}}_s(c_{17}) + \widehat{\mathrm{P}}_s(c_{18}) + \widehat{\mathrm{P}}_s(c_{16}) = \widehat{\mathrm{P}}_s(c_2) \\
\widehat{\mathrm{P}}_s(c_8) + \widehat{\mathrm{P}}_s(c_9) + \widehat{\mathrm{P}}_s(c_{10}) = \widehat{\mathrm{P}}_s(c_3) \\
\widehat{\mathrm{P}}_s(c_{17}) + \widehat{\mathrm{P}}_s(c_{18}) + \widehat{\mathrm{P}}_s(c_{16}) = \widehat{\mathrm{P}}_s(c_7) \\
\widehat{\mathrm{P}}_s(c_{17}) + \widehat{\mathrm{P}}_s(c_{18}) = \widehat{\mathrm{P}}_s(c_{15})
\end{cases}$$

La solution (cf. figure 6.12) telle que les probabilités des concepts soient dans l'intervalle $[0;1]$ est la suivante :

$$\begin{cases}
\widehat{\mathrm{P}}_s(c_{13}) = 17 - 12 \cdot \sqrt{2} & \widehat{\mathrm{P}}_s(c_{14}) = 10 \cdot \sqrt{2} - 14 \\
\widehat{\mathrm{P}}_s(c_{11}) = 5 \cdot \sqrt{2} - 7 & \widehat{\mathrm{P}}_s(c_4) = 3 - 2 \cdot \sqrt{2} \\
\widehat{\mathrm{P}}_s(c_{16}) = \dfrac{3}{2} - \sqrt{2} & \widehat{\mathrm{P}}_s(c_{17}) = \dfrac{3}{4} - \dfrac{\sqrt{2}}{2} \\
\widehat{\mathrm{P}}_s(c_8) = 2 - \dfrac{4 \cdot \sqrt{2}}{3} & \widehat{\mathrm{P}}_s(c_{12}) = 5 \cdot \sqrt{2} - 7 \\
\widehat{\mathrm{P}}_s(c_{18}) = \dfrac{3}{4} - \dfrac{\sqrt{2}}{2} & \widehat{\mathrm{P}}_s(c_9) = 2 - \dfrac{4 \cdot \sqrt{2}}{3} \\
\widehat{\mathrm{P}}_s(c_{10}) = 2 - \dfrac{4 \cdot \sqrt{2}}{3} & \widehat{\mathrm{P}}_s(c_5) = 3 - 2 \cdot \sqrt{2} \\
\widehat{\mathrm{P}}_s(c_6) = 3 - 2 \cdot \sqrt{2} & \widehat{\mathrm{P}}_s(c_1) = 6 - 4 \cdot \sqrt{2} \\
\widehat{\mathrm{P}}_s(c_2) = 6 - 4 \cdot \sqrt{2} & \widehat{\mathrm{P}}_s(c_3) = 6 - 4 \cdot \sqrt{2} \\
\widehat{\mathrm{P}}_s(c_7) = 3 - 2 \cdot \sqrt{2} & \widehat{\mathrm{P}}_s(c_{15}) = \dfrac{3}{2} - \sqrt{2}
\end{cases}$$

Cependant, il n'y a parfois aucune solution acceptable tel qu'à chaque concept est attribué une probabilité dans l'intervalle $[0;1]$. La hiérarchie de la figure 6.13 montre l'approximation aberrante qui répond aux contraintes imposées avec $\rho = 1$ et $\epsilon = 0$. Cela met en évidence la nécessité soit de réévaluer le paramètre ρ soit de relaxer la contrainte de complétude en ajoutant $\epsilon \neq 0$ fils supplémentaires à chaque concept noeud.

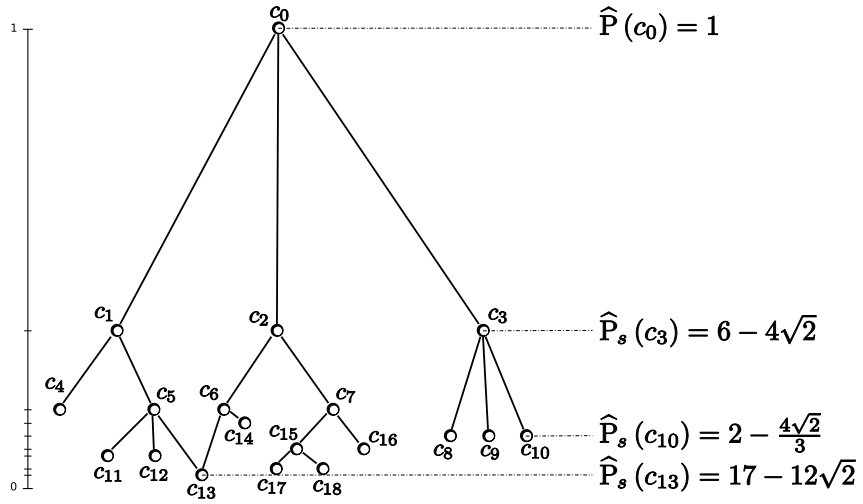


FIG. 6.12 – Application de l'approximation $\widehat{\mathrm{P}}_s$ avec $\widehat{\mathrm{P}}(c_0) = 1$, $\rho = 1$ et $\epsilon = 0$

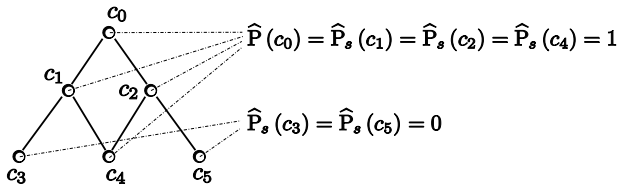


FIG. 6.13 – Un cas aberrant de l'approximation $\widehat{\mathrm{P}}_s$ avec $\widehat{\mathrm{P}}(c_0) = 1$, $\rho = 1$ et $\epsilon = 0$

A l'échelle d'une ontologie réelle, la résolution du système d'équations engendré reste néanmoins un problème difficile à résoudre. La difficulté réside dans la détermination de la probabilté associée aux concepts ayant plusieurs parents qui donne lieu à des équations non linéaires dont le degré est de l'ordre du nombre de parents dont ils héritent. Pour résoudre ce problème, nous détaillons maintenant un algorithme efficace qui offre une convergence rapide avec une précision maîtrisée.

Principe de l'algorithme

L'algorithme propose d'approcher la solution du système par valeurs inférieures. L'initialisation des probabilités recherchées est conduite avec l'hypothèse d'une disjonction systématique. Ensuite, chaque itération permet un réajustement des probabilités vers la solution recherchée. La probabilité associée à chaque fils non exclusif sous hypothèse de disjonction est nulle. Lors de la première itération, une borne inférieure de chacune de ces probabilités est évaluée, ce qui remet en cause la disjonction. Les probabilités associées aux autres concepts sur la base d'une disjonction sont alors mises à jour en conséquence.

Une nouvelle itération donne lieu à une nouvelle borne inférieure pour chaque probabilité associée à un fils non exclusif ce qui entraîne à nouveau un recalcul des probabilités associées aux autres concepts et ainsi de suite.

La solution sous l'hypothèse d'une disjonction systématique étant relativement proche de la solution recherchée, cet algorithme converge très rapidement vers la solution que l'on recherche. Il est nécessaire de paramétrer la précision désirée en fixant le nombre de décimales correctes attendues.

Avant de présenter l'algorithme général, nous proposons d'en expliciter le principe de manière plus détaillée en nous appuyant sur notre exemple. Avant toute chose, nous calculons les probabilités minimales associées à chaque concept en nous basant sur l'hypothèse d'une disjonction systématique (cf. figure 6.14). Nous partons donc de la racine à laquelle est affectée la probabilité maximale que l'on divise en trois pour chacun de ses fils et ainsi de suite. Du fait de la disjonction supposée a priori, on retrouve une probabilité nulle associée au concept c_{13}. Le concept c_{13} ne peut pas avoir une probabilité associée plus élevée que celle d'un de ses frères :

$$\widehat{\mathrm{P}}_s\left(c_{13}\right) < \min\left\{\frac{\widehat{\mathrm{P}}_s(c_5)}{3}; \frac{\widehat{\mathrm{P}}_s(c_6)}{2}\right\} = \frac{1}{18} \approx 0.0556$$

Cela nous permet alors de fixer la probabilité minimale de chacun des frères de c_{13} :

$$\widehat{\mathrm{P}}_s\left(c_{11}\right) = \widehat{\mathrm{P}}_s\left(c_{12}\right) = \frac{\widehat{\mathrm{P}}_s(c_5) - \frac{1}{18}}{2} = \frac{1}{18} \approx 0.0556$$

$$\widehat{\mathrm{P}}_s\left(c_{14}\right) = \widehat{\mathrm{P}}_s\left(c_6\right) - \frac{1}{18} = \frac{1}{9} \approx 0.1111$$

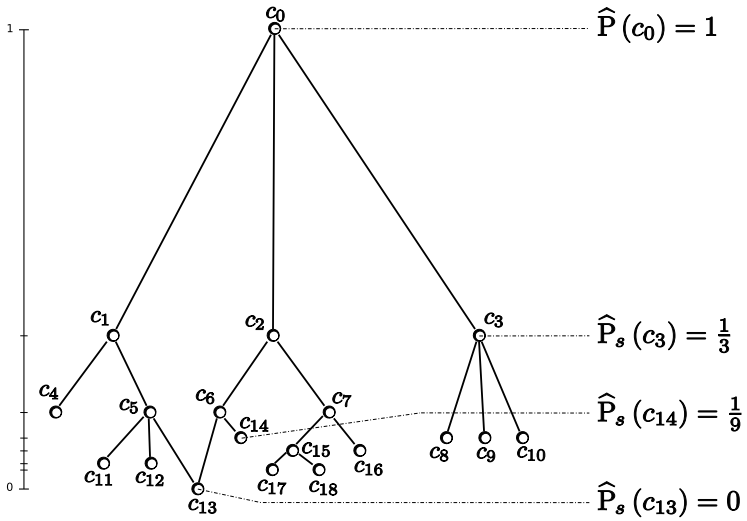


FIG. 6.14 – Calcul des probabilités $\widehat{\mathrm{P}}_s$ sous hypothèse de disjonction

La difficulté de la recherche de la solution vient du fait que la probabilité $\widehat{\mathrm{P}}_s\left(c_1\right) = \widehat{\mathrm{P}}_s\left(c_2\right) = \widehat{\mathrm{P}}_s\left(c_3\right)$ associée aux fils de la racine dépend de la probabilité

$\widehat{\mathrm{P}}_s(c_{13})$ associée à c_{13} (cf. figure 6.15). Or, la probabilité de c_{13} dépend des probabilités associées à c_1 et c_2 (cf. figure 6.16). Le principe de l'algorithme est de faire une première approximation sur la base des probabilités obtenues avec l'hypothèse de disjonction ce qui nous donne une borne inférieure pour chaque concept comme le présente la figure 6.14. Cela nous permet de calculer une première borne inférieure non nulle pour c_{13} :

$$\widehat{\mathrm{P}}_s(c_{13}) = \widehat{\mathrm{P}}_s(c_5) \cdot \widehat{\mathrm{P}}_s(c_6) = \tfrac{1}{36} \approx 0.0278$$

Cette borne inférieure pour c_{13} va nous permettre de remettre en cause la disjonction (cf. figure 6.15). Parmi les instances de la racine, les instances de c_{13} sont dupliquées pour être réparties dans les concepts c_1 et c_2, c'est pourquoi on doit ajouter $\widehat{\mathrm{P}}_s(c_{13})$ à $\widehat{\mathrm{P}}_s(c_0)$ avant de répartir équitablement les instances dans les trois fils de la racine :

$$\widehat{\mathrm{P}}_s(c_1) = \widehat{\mathrm{P}}_s(c_2) = \widehat{\mathrm{P}}_s(c_3) = \tfrac{\widehat{\mathrm{P}}(c_0)+\widehat{\mathrm{P}}_s(c_{13})}{3} = \tfrac{37}{108} \approx 0.3426$$

Nous répartissons ensuite les instances de c_1, c_2 et c_3 dans leurs fils respectifs. Nous obtenons entre autre $\widehat{\mathrm{P}}_s(c_5) = \widehat{\mathrm{P}}_s(c_6) = \tfrac{37}{216} \approx 0.1713$. Ces nouvelles approximations nous permettent de recalculer une nouvelle borne inférieure plus élévée pour la probabilité associée à c_{13} : $\widehat{\mathrm{P}}_s(c_{13}) = \tfrac{1369}{46656} \approx 0.0293$.

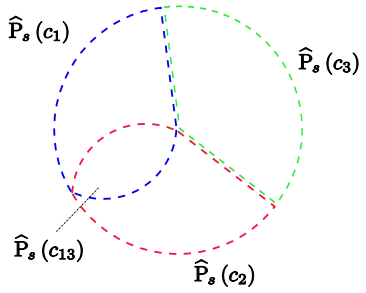


FIG. 6.15 – Dépendance de $\widehat{\mathrm{P}}_s(c_1)$, $\widehat{\mathrm{P}}_s(c_2)$ et $\widehat{\mathrm{P}}_s(c_3)$ vis-à-vis de $\widehat{\mathrm{P}}_s(c_{13})$

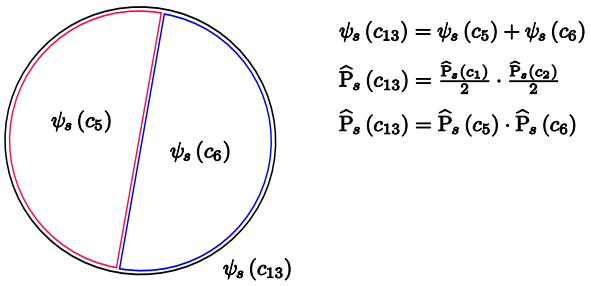


FIG. 6.16 – Dépendance de $\widehat{\mathrm{P}}_s(c_{13})$ vis-à-vis de $\widehat{\mathrm{P}}_s(c_1)$ et $\widehat{\mathrm{P}}_s(c_2)$

On réitère cela autant de fois que nécessaire pour atteindre la précision requise. Pour terminer, on réévalue les probabilités des frères de c_{13} de manière

à avoir une répartition complète des instances de leur père. Le tableau 6.2 trace les résultats de l'algorithme après la phase d'initialisation ainsi qu'après chaque itération remettant en cause la disjonction initiale et enfin après l'ajustement des frères de c_{13}. Après la 6ème itération seulement, nous obtenons 7 décimales fixes.

	$\widehat{\mathrm{P}}_s$ initial	$\widehat{\mathrm{P}}_s$ à l'itération				$\widehat{\mathrm{P}}_s$ final
		1	2	...	6,7,...	
c_0	1	1	1		1	1
c_1	0,3333333	0,3425926	0,3431141		0,3431458	0,3431458
c_2	0,3333333	0,3425926	0,3431141		0,3431458	0,3431458
c_3	0,3333333	0,3425926	0,3431141		0,3431458	0,3431458
c_4	0,1666667	0,1712963	0,1715571		0,1715729	0,1715729
c_5	0,1666667	0,1712963	0,1715571		0,1715729	0,1715729
c_6	0,1666667	0,1712963	0,1715571		0,1715729	0,1715729
c_7	0,1666667	0,1712963	0,1715571		0,1715729	0,1715729
c_8	0,1111111	0,1141975	0,1143714		0,1143819	0,1143819
c_9	0,1111111	0,1141975	0,1143714		0,1143819	0,1143819
c_{10}	0,1111111	0,1141975	0,1143714		0,1143819	0,1143819
c_{11}	0,0555555	0,0570988	0,0571857		0,0710678	0,0710678
c_{12}	0,0555555	0,0570988	0,0571857		0,0710678	0,0710678
c_{13}	0	0,0277777	0,0293424		0,0294373	0,0294373
c_{14}	0,1111111	0,1141975	0,1143714		0,1143819	0,1421356
c_{15}	0,0833333	0,08564815	0,0857785		0,0857864	0,0857864
c_{16}	0,0833333	0,08564815	0,0857785		0,0857864	0,0857864
c_{17}	0,0416666	0,04282407	0,0428893		0,0428932	0,0428932
c_{18}	0,0416666	0,04282407	0,0428893		0,0428932	0,0428932

TAB. 6.2 – Trace de l'algorithme sur notre exemple

L'algorithme 6.1 explicite l'algorithme général permettant de fournir la solution sur une hiérarchie de subsomption quelconque. Nous pouvons distinguer quatre grandes étapes dans cet algorithme :

1. La première étape de l'algorithme (Initialisation de $\widehat{\mathrm{P}}_s$ sous hypothèse de disjonction systématique) consiste à initialiser les probabilités recherchées avec une borne inférieure la plus élevée possible. L'algorithme proposé est donc fondé sur un parcours de la hiérarchie lors duquel la probabilité $\widehat{\mathrm{P}}_s$ est initialisée sous hypothèse de disjonction systématique. On répartit de manière uniforme et disjointe la totalité des instances d'un père vers les extensions de ses fils lorsque tous ses fils sont des fils exclusifs. Dans le cas ou un ou plusieurs fils sont des fils non exclusifs, leur extension est initialement considérée comme vide et sera réajustée par la suite. Les fils exclusifs qui sont les frères de ces fils non exclusifs doivent alors se partager la totalité des instances de leur père de manière disjointe et uniforme. Mais le fait qu'ils aient des frères dont la probabilité associée va être revue à la hausse par la remise en cause de la disjonction au niveau d'un de leur subsumant, leur probabilité serait revue à la baisse. Pour garantir la convergence de l'algorithme vers la solution, il nous faut approcher ces probabilités par valeur inférieure. Nous allons donc attribuer une valeur

maximale inférieure à la probabilité recherchée qui sera réajustée en temps voulu.

2. La seconde étape (Calcul d'une nouvelle borne inférieure pour les concepts multi-héritants) se base sur les probabilités associées aux concepts pour recalculer celles associées aux concepts multi-héritants. Pour cela, on utilise la formule du contenu informationnel généralisé sur l'ensemble des parents de chaque concept multi-héritant. On ajoute à cela une certaine quantité d'information fixée à l'aide du paramètre ρ.

3. La troisième étape (Calcul d'une nouvelle borne inférieure pour les autres concepts) consiste à revoir la probabilité des concepts non multi-héritants en tenant compte de la non disjonction dont l'importance est donnée par les probabilités associées aux concepts multi-héritants qui ont été calculées à l'étape précédente.

4. La dernière étape (Recalcul d'une nouvelle borne inférieure pour les autres concepts) est similaire à la précédente sauf que cette fois-ci, les probabilités des concepts multi-héritants ayant atteint un palier de stabilité, on peut réajuster les probabilité de leurs frères à la hausse de manière à garantir la complétude.

L'algorithme converge très rapidemment vers une solution stable pour une précision donnée, c'est-à-dire que nous avons un certain nombre de décimales fixé qui n'évolue plus.

Algorithme 6.1 : Algorithme général pour la détermination de $\widehat{\mathrm{P}}_s$

Entrées : – $\mathcal{H}$, hiérarchie de subsomption,
– $\widehat{\mathrm{P}}(c_0)$, Probabilité de la racine
Sorties : – $\widehat{\mathrm{P}}_s$, relation qui associe à un concept c_i sa probabilité

`// Initialisation de` $\widehat{\mathrm{P}}_s$ `sous hypothèse disjonctive`
1 $\widehat{\mathrm{P}}_s = \text{initialiser}(\mathcal{H}, \widehat{\mathrm{P}}(c_0))$;
2 **répéter**
3 **répéter**
`// Calcul d'une nouvelle borne inférieure pour les concepts multi-héritants`
4 $\widehat{\mathrm{P}}_s = \text{calculer1}(\mathcal{H}, \widehat{\mathrm{P}}(c_0), \widehat{\mathrm{P}}_s)$;
`// Calcul d'une nouvelle borne inférieure pour les autres concepts`
5 $\widehat{\mathrm{P}}_s = \text{calculer2}(\mathcal{H}, \widehat{\mathrm{P}}(c_0), \widehat{\mathrm{P}}_s)$;
jusqu'à $\forall c_i \in \mathcal{C}, \widehat{\mathrm{P}}_s(c_i)$ *stable*
`// Réajustement des probabilités des autres concepts`
6 $\widehat{\mathrm{P}}_s = \text{reajuster}(\mathcal{H}, \widehat{\mathrm{P}}(c_0), \widehat{\mathrm{P}}_s)$;
jusqu'à $\forall c_i \in c_0^{\sqsupset\natural}, \widehat{\mathrm{P}}_s(c_i)$ *stable*
7 **retourner** $\widehat{\mathrm{P}}_s$;

Algorithme 6.2 : Initialisation de $\widehat{\mathrm{P}}_s$ sous hypothèse de disjonction systématique (initialiser)

Entrées : – $\mathcal{H}$, hiérarchie de subsomption,
– $\widehat{\mathrm{P}}(c_0)$, Probabilité de la racine
Sorties : – $\widehat{\mathrm{P}}_s$, relation qui associe à un concept c_i sa probabilité

$\widehat{\mathrm{P}}_s(c_0) = \widehat{\mathrm{P}}(c_0)$;
pour chaque $c_i \in c_0^{\sqsupset}$ **faire**
  si $|c_i^{\prec}| = 1$ **alors**
    si $\left|(c_i^{\star})^{\succ}\right| = \left|(c_i^{\star})^{\succ !}\right|$ **alors**
      $\widehat{\mathrm{P}}_s(c_i) = \frac{\widehat{\mathrm{P}}_s(c_i^{\star})}{\left|(c_i^{\star})^{\succ}\right|}$;
    sinon
      $$\widehat{\mathrm{P}}_s(c_i) = \frac{\widehat{\mathrm{P}}_s(c_i^{\star}) - \sum_{c_x \in (c_i^{\star})^{\succ !}} \min_{c_y \in c_x^{\prec}} \left\{ \frac{\widehat{\mathrm{P}}_s(c_y)}{\left|c_y^{\succ}\right|} \right\}}{\left|(c_i^{\star})^{\succ !}\right|} ;$$
  sinon
    $\widehat{\mathrm{P}}_s(c_i) = 0$;
retourner $\widehat{\mathrm{P}}_s$;

Algorithme 6.3 : Calcul d'une nouvelle borne inférieure pour les concepts multi-héritants (calculer1)

Entrées : – $\mathcal{H}$, hiérarchie de subsomption,
– $\widehat{\mathrm{P}}(c_0)$, Probabilité de la racine,
– $\widehat{\mathrm{P}}_s$, relation qui associe à un concept c_i sa probabilité
Sorties : – $\widehat{\mathrm{P}}_s$, relation qui associe à un concept c_i sa probabilité

`// Calcul d'une nouvelle borne inférieure pour les concepts multi-héritants`
pour chaque $c_i \in c_0^{\sqsupset !}$ **faire**
  $$\psi_s^{\cup}(c_i^{\prec}) = \sum_{k=1}^{|c_i^{\prec}|} \left((-1)^{k+1} \cdot \sum_{\mathcal{C}_x \in \mathcal{P}_k(c_i^{\prec})} \psi^{\cap}(\mathcal{C}_x) \right) ;$$
  $\widehat{\mathrm{P}}_s(c_i) = a^{-\left[\psi_s^{\cup}(c_i^{\prec}) + \log_a r\right]}$;
retourner $\widehat{\mathrm{P}}_s$;

Algorithme 6.4 : Calcul d'une nouvelle borne inférieure pour les autres concepts (calculer2)

Entrées : – $\mathcal{H}$, hiérarchie de subsomption,
– $\widehat{\mathrm{P}}(c_0)$, Probabilité de la racine,
– $\widehat{\mathrm{P}}_s$, relation qui associe à un concept c_i sa probabilité
Sorties : – $\widehat{\mathrm{P}}_s$, relation qui associe à un concept c_i sa probabilité

`// Calcul d'une nouvelle borne inférieure pour les autres concepts`

pour chaque $c_i \in c_0^{\sqsupset!}$ **faire**

  si $\left|(c_i^\star)^{\succ}\right| = \left|(c_i^\star)^{\succ!}\right|$ **alors**

$$\widehat{\mathrm{P}}_s(c_i) = \frac{\widehat{\mathrm{P}}_s(c_i^\star) - \sum_{k=2}^{\left|(c_i^\star)^{\succ}\right|} \left((-1)^{k+1} \cdot \sum_{\substack{c_y \in \cap_{\mathcal{C}_x}^{\sqsupset} \wedge \\ \mathcal{C}_x \in \mathcal{P}_k(c_i^{\succ}) \wedge \\ \cap_{\mathcal{C}_x}^{\sqsupset} \cap c_y^{\sqsubset} = \emptyset}} \widehat{\mathrm{P}}_s(c_y) \right)}{\left|(c_i^\star)^{\succ}\right|} ;$$

  sinon

$$\widehat{\mathrm{P}}_s(c_i) = \frac{\widehat{\mathrm{P}}_s(c_i^\star) - \sum_{c_x \in (c_i^\star)^{\succ!}} \min_{c_y \in c_x^{\prec}} \left\{ \frac{\widehat{\mathrm{P}}_s(c_y)}{\left|c_y^{\succ}\right|} \right\} - \sum_{k=2}^{\left|(c_i^\star)^{\succ}\right|} \left((-1)^{k+1} \cdot \sum_{\substack{c_y \in \cap_{\mathcal{C}_x}^{\sqsupset} \wedge \\ \mathcal{C}_x \in \mathcal{P}_k(c_i^{\succ}) \wedge \\ \cap_{\mathcal{C}_x}^{\sqsupset} \cap c_y^{\sqsubset} = \emptyset}} \widehat{\mathrm{P}}_s(c_y) \right)}{\left|(c_i^\star)^{\succ!}\right|} ;$$

retourner $\widehat{\mathrm{P}}_s$;

Algorithme 6.5 : Réajustement des probabilités des autres concepts (reajuster)

Entrées : – $\mathcal{H}$, hiérarchie de subsomption,
– $\widehat{\mathrm{P}}(c_0)$, Probabilité de la racine,
– $\widehat{\mathrm{P}}_s$, relation qui associe à un concept c_i sa probabilité

Sorties : – $\widehat{\mathrm{P}}_s$, relation qui associe à un concept c_i sa probabilité

`// Réajustement des probabilités des autres concepts`

pour chaque $c_i \in c_0^{\sqsupset !}$ **faire**

si $\left|(c_i^\star)^{\succ}\right| = \left|(c_i^\star)^{\succ !}\right|$ **alors**

$$\widehat{\mathrm{P}}_s(c_i) = \frac{\widehat{\mathrm{P}}_s(c_i^\star) - \sum_{k=2}^{\left|(c_i^\star)^{\succ}\right|} \left((-1)^{k+1} \cdot \sum_{\substack{c_y \in \cap_{\overline{\mathcal{C}}_x}^{\sqsupset} \wedge \\ \mathcal{C}_x \in \mathcal{P}_k(c_i^{\succ}) \wedge \\ \cap_{\overline{\mathcal{C}}_x}^{\sqsupset} \cap c_y^{\sqsubset} = \emptyset}} \widehat{\mathrm{P}}_s(c_y) \right)}{\left|(c_i^\star)^{\succ}\right|} ;$$

sinon

$$\widehat{\mathrm{P}}_s(c_i) = \frac{\widehat{\mathrm{P}}_s(c_i^\star) - \sum_{c_x \in (c_i^\star)^{\succ !}} \widehat{\mathrm{P}}_s(c_x) - \sum_{k=2}^{\left|(c_i^\star)^{\succ}\right|} \left((-1)^{k+1} \cdot \sum_{\substack{c_y \in \cap_{\overline{\mathcal{C}}_x}^{\sqsupset} \wedge \\ \mathcal{C}_x \in \mathcal{P}_k(c_i^{\succ}) \wedge \\ \cap_{\overline{\mathcal{C}}_x}^{\sqsupset} \cap c_y^{\sqsubset} = \emptyset}} \widehat{\mathrm{P}}_s(c_y) \right)}{\left|(c_i^\star)^{\succ !}\right|} ;$$

retourner $\widehat{\mathrm{P}}_s$;

6.5 Généralisation de l'analogie

La généralisation du contenu informationnel partagé présenté dans ce chapitre laisse inchangée l'analogie proposée au chapitre 5. Cependant, le contenu informationnel global d'un sous-ensemble de concepts nous permet d'exprimer le contenu informationnel de l'ensemble des concepts de la hiérarchie $\psi^{\cup}(\mathcal{C})$. Ce contenu informationnel est analogue à la quantité n qui désigne l'importance de l'intension $\mathcal{I}$ de l'ensemble des concepts de la hiérarchie (cf. tableau 6.3).

$(c_i, c_j) \in \mathcal{C}^2$	
n_i	$\psi(c_i)$
n_j	$\psi(c_j)$
n_{ij}	$\psi^{\cap}(\{c_i, c_j\})$
n	$\psi^{\cup}(\mathcal{C})$

TAB. 6.3 – Analogie entre l'importance d'une description intensionnelle et le contenu informationnel

6.5.1 Retour sur les mesures existantes

Avec la généralisation du contenu informationnel partagé, les réécritures des mesures existantes proposées au chapitre 5 sont adaptées pour une hiérarchie avec héritage multiple. Cependant, certaines des mesures existantes traitent l'héritage mulitple de manière différente.

Concernant la mesure de Stojanovic, la réécriture proposée est équivalente face à la prise en compte de l'héritage multiple. La mesure de Wu & Palmer tout comme la proportion de spécificité partagée se limite initialement à un arbre ; leur réécriture constitue donc une généralisation à une hiérarchie.

Les mesures de Resnik, Jiang & Conrath et Lin utilisent le contenu informationnel du subsumant commun le plus spécifique $\psi(c_{ij})$ qui correspond, dans un arbre de subsomption, au contenu informationnel partagé $\psi^{\cap}(\{c_i, c_j\})$. Dans une hiérarchie, le contenu informationnel partagé peut correspondre à celui d'un sous-ensemble de concepts et non plus à celui d'un seul et unique concept.

En considérant le plus court chemin, Rada ne tient pas non plus compte de la multiplicité des subsumants communs qui participent au contenu informationnel partagé. On remarque également que deux concepts ayant un fils en commun sont donc séparés par un chemin très court tandis qu'ils n'ont peut être que la racine comme subsumant commun.

La mesure de Zhong utilise la notion de longueur du plus long chemin du concept considéré jusqu'à la racine. Tout comme les mesures de Rada, Resnik, Jiang & Conrath et Lin celle de Zhong laisse donc de côté une partie de l'information que renferme le contenu informationnel partagé.

De manière générale, l'utilisation du contenu informationnel partagé propose une généralisation à une hiérarchie plus adaptée en tenant compte de toute l'information mise en jeu.

Remarque. On remarque que les mesures sémantiques de la littérature ne tiennent jamais compte de la quantité d'information globale de la hiérarchie $\psi^{\cup}(\mathcal{C})$ qui est analogue au paramètre n. Cette quantité d'information globale est très grande au regard des quantités $\psi(c_i)$, $\psi(c_j)$ et $\psi^{\cap}(\{c_i, c_j\})$ ce qui rend notamment les indices d'écart à l'indépendance difficilement utilisables.

6.5.2 Propriétés métriques et ordinales

Les propriétés qui ne sont pas respectées dans un arbre de subsomption ne le sont pas non plus dans le cadre moins contraint d'une hiérarchie de subsomption. Cependant, l'inégalité de Maguitman qui est respectée pour certaines valeurs de α et θ ne l'est plus lorsque l'héritage multiple est permis.

Proposition 6.1 *Dans une hiérarchie de subsomption, les similarités $\widetilde{\sigma}_\alpha$ ne respectent pas l'inégalité de Maguitman.*

Preuve.

Prenons un contre-exemple (cf. figure 6.17) avec trois concepts : c_1, c_2, c_3 tel que c_1 et c_2 subsument c_3, $\psi(c_1) = \psi(c_2) = 1$, $\psi(c_3) = 2$ et $\psi(c_{12}) = 0$.

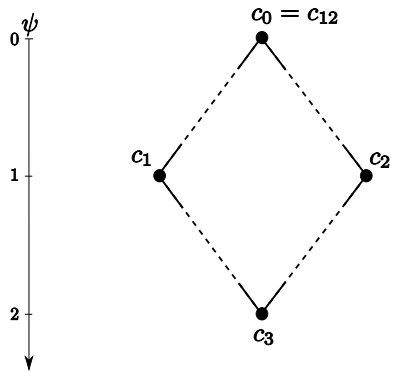


FIG. 6.17 – Contre-exemple concernant le respect de l'inégalité de Maguitman par les similarités $\widetilde{\sigma}_\alpha$

On calcule les similarités $\widetilde{\sigma}_\alpha$ sur cet exemple :

$$\widetilde{\sigma}_\alpha(c_1, c_2) = \frac{\psi(c_{12})}{\mu_\alpha\big(\psi(c_1), \psi(c_2)\big)} = 0$$
$$\widetilde{\sigma}_\alpha(c_1, c_3) = \widetilde{\sigma}_\alpha(c_3, c_2) > \frac{1}{2} > 0$$

On en déduit,

$$\widetilde{\sigma}_\alpha(c_1, c_2) < \widetilde{\sigma}_\alpha(c_1, c_3) \cdot \widetilde{\sigma}_\alpha(c_3, c_2)$$

Donc l'inégalité de Maguitman n'est pas respectée par les similarités $\widetilde{\sigma}_\alpha$ quelque soit la valeur de α.

Proposition 6.2 *Les similarités $\widetilde{\sigma}_\theta$ ne respectent pas l'inégalité de Maguitman.*

Preuve.

Prenons un contre-exemple (cf. figure 6.18) avec trois concepts : c_1, c_2, c_3 tel que c_1 et c_2 subsument c_3, $\psi(c_1) = \psi(c_2) = 1$, $\psi(c_3) = 2$ et $\psi(c_{12}) = 0$.

On calcule les θ-similarités sur cet exemple :

$$\widetilde{\sigma}_\theta(c_1, c_2) = \frac{\theta \cdot (\psi(c_{12}))}{\psi(c_1) + \psi(c_2) + (\theta - 2) \cdot \psi(c_{12})} = 0$$

$$\widetilde{\sigma}_\alpha(c_1, c_3) = \widetilde{\sigma}_\alpha(c_3, c_2) > \frac{\theta \cdot 1}{1 + 2 + (\theta - 2) \cdot 1} > 0$$

On en déduit,

$$\widetilde{\sigma}_\theta(c_1, c_2) < \widetilde{\sigma}_\theta(c_1, c_3) \cdot \widetilde{\sigma}_\theta(c_3, c_2)$$

Donc l'inégalité de Maguitman n'est pas respectée par les similarités $\widetilde{\sigma}_\theta$ quelque soit la valeur de θ.

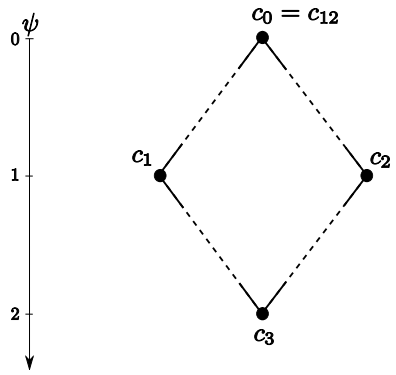


FIG. 6.18 – Contre-exemple concernant le respect de l'inégalité de Maguitman par les similarités $\widetilde{\sigma}_\theta$

6.5.3 Ressemblance entre deux sous-ensembles de concepts

Nous terminons ce chapitre en considérant le cas où une mesure sémantique entre sous-ensembles de concepts est nécessaire. C'est le cas typique dans lequel des documents sont indexés par un ensemble de mots-clés qui référencent des concepts d'une ontologie, l'objectif étant de retrouver la liste des documents les plus pertinents pour une requête donnée. Pour ce faire, on cherche à évaluer la ressemblance entre l'ensemble des mots-clés de la requête et l'ensemble des mots-clés qui indexent les documents. Une mesure composite qui consiste à agréger des similarités deux à deux n'est pas toujours souhaitable puisque deux concepts très proches qui indexent un même document vont faire augmenter de manière non justifiée la pertinence du document en question.

Une solution plus adaptée est de considérer le contenu informationnel partagée par deux sous-ensembles de concepts $\mathcal{C}_i$ et $\mathcal{C}_j$ qui peut se calculer comme suit :

$$\psi^{\cap}\left(\mathcal{C}_i, \mathcal{C}_j\right) = \psi^{\cup}\left(\Gamma\left(\cup_{\mathcal{C}_i}^{\sqsubseteq} \cap \cup_{\mathcal{C}_j}^{\sqsubseteq}\right)\right) \tag{6.13}$$

Nous avons vu au chapitre 6 comment calculer le contenu informationnel d'un sous-ensemble de concepts (contenu informationnel global $\psi^{\cup}$) de telle sorte que l'analogie proposée peut être adaptée comme suit :

$(c_i, c_j) \in \mathcal{C}^2$	
n_i	$\psi^{\cup}\left(\mathcal{C}_i\right)$
n_j	$\psi^{\cup}\left(\mathcal{C}_j\right)$
n_{ij}	$\psi^{\cap}\left(\mathcal{C}_i, \mathcal{C}_j\right)$
n	$\psi^{\cup}\left(\mathcal{C}\right)$

TAB. 6.4 – Adaptation de l'analogie entre l'importance d'une description intensionnelle et le contenu informationnel pour deux sous-ensembles de concepts

Comme nous l'avons fait dans la section 5.3, les diverses mesures présentées au chapitre 3 peuvent être adaptées pour évaluer une liaison entre deux sous-ensembles de concepts. Nous reprenons simplement les familles σ_θ et σ_α :

$$\widetilde{\sigma}'_\theta\Big(\mathcal{C}_i, \mathcal{C}_j\Big) = \left\{\frac{\theta \cdot \psi^{\cap}\left(\mathcal{C}_i, \mathcal{C}_j\right)}{(\theta-2)\cdot\psi^{\cap}\left(\mathcal{C}_i, \mathcal{C}_j\right) + \psi^{\cup}\left(\mathcal{C}_i\right) + \psi^{\cup}\left(\mathcal{C}_j\right)}\right\}_{\theta\in\mathbb{R}_+^*}$$

$$\widetilde{\sigma}'_\alpha\Big(\mathcal{C}_i, \mathcal{C}_j\Big) = \left\{\frac{\psi^{\cap}\left(\mathcal{C}_i, \mathcal{C}_j\right)}{\mu_\alpha\left(\psi^{\cup}\left(\mathcal{C}_i\right), \psi^{\cup}\left(\mathcal{C}_j\right)\right)}\right\}_{\alpha\in\mathbb{R}} \text{avec, } \mu_\alpha = \left(\frac{\psi^{\cup}\left(\mathcal{C}_i\right)^\alpha + \psi^{\cup}\left(\mathcal{C}_j\right)^\alpha}{2}\right)^{\frac{1}{\alpha}}$$

Tandis que les démonstrations concernant la préordonnance restent valables, il n'en est pas de même pour l'inégalité de Maguitman.

Proposition 6.3 *Que la hiérarchie de subsomption soit restreinte ou non à un arbre, les similarités $\widetilde{\sigma}'_\theta$ et $\widetilde{\sigma}'_\alpha$ ne respectent pas l'inégalité de Maguitman.*

Preuve.

Prenons un contre-exemple (cf. figure 6.19) avec trois sous-ensembles de concepts : $\{c_1\}$, $\{c_2\}$ et $\{c_3, c_4\}$ tel que $\psi\left(c_{12}\right) = \psi\left(c_{14}\right) = \psi\left(c_{23}\right) = 0$ et $\psi\left(c_{13}\right) = \psi\left(c_{24}\right) = 1$.

On calcule les α-similarités sur cet exemple :

$$\widetilde{\sigma}'_\alpha\Big(\{c_1\}, \{c_2\}\Big) = \frac{\psi^{\cap}\left(\{c_1\}, \{c_2\}\right)}{\mu_\alpha\left(\psi^{\cup}\left(\{c_1\}\right), \psi^{\cup}\left(\{c_2\}\right)\right)} = 0$$

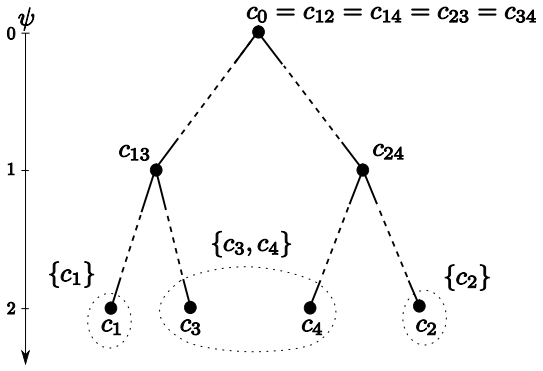


FIG. 6.19 – Contre-exemple concernant le respect de l'inégalité de Maguitman des similarités $\widetilde{\sigma}'_\theta$ et $\widetilde{\sigma}'_\alpha$

$$\widetilde{\sigma}'_\alpha\Big(\{c_1\},\{c_3,c_4\}\Big) = \widetilde{\sigma}'_\alpha\Big(\{c_3,c_4\},\{c_2\}\Big) > 0$$

On en déduit,

$$\widetilde{\sigma}'_\alpha\Big(\{c_1\},\{c_2\}\Big) < \widetilde{\sigma}'_\alpha\Big(\{c_1\},\{c_3,c_4\}\Big) \cdot \widetilde{\sigma}'_\alpha\Big(\{c_3,c_4\},\{c_2\}\Big)$$

Donc l'inégalité de Maguitman n'est pas respectée par les similarités $\widetilde{\sigma}'_\alpha$ quelque soit la valeur de α.

On calcule les θ-similarités sur cet exemple :

$$\widetilde{\sigma}'_\theta\Big(\{c_1\},\{c_2\}\Big) = \frac{\theta \cdot (\psi^\cap(\{c_1\},\{c_2\}))}{\psi^\cup(\{c_1\}) + \psi^\cup(\{c_1\}) + (\theta-2)\cdot\psi^\cap(\{c_1\},\{c_2\})} = 0$$

$$\widetilde{\sigma}'_\alpha\Big(\{c_1\},\{c_3,c_4\}\Big) = \widetilde{\sigma}'_\alpha\Big(\{c_3,c_4\},\{c_2\}\Big) > 0$$

On en déduit,

$$\widetilde{\sigma}'_\theta\Big(\{c_1\},\{c_2\}\Big) < \widetilde{\sigma}'_\theta\Big(\{c_1\},\{c_3,c_4\}\Big) \cdot \widetilde{\sigma}'_\theta\Big(\{c_3,c_4\},\{c_2\}\Big)$$

Donc l'inégalité de Maguitman n'est pas respectée par les similarités $\widetilde{\sigma}'_\theta$ quelque soit la valeur de θ.

6.6 Conclusion

Nous avons consacré ce chapitre à la généralisation de nos propositions pour prendre en compte l'héritage multiple. Nous avons approfondi la notion de contenu informationnel en proposant le contenu informationnel global $\psi^\cup$ qui est une adaptation pour un sous-ensemble de concepts d'une hiérarchie de subsomption. De la même manière, nous avons également adapté le contenu informationnel partagé $\psi^\cap$.

Nous avons lorsque cela était nécessaire adapté les approximations du chapitre 4 sur lesquelles repose le contenu informationnel. Seules les approximations $\widehat{\mathrm{P}}_p$ et $\widehat{\mathrm{P}}_s$ relevant d'une approche descendante ont dû être adaptées. L'approximation $\widehat{\mathrm{P}}_s$ qui considère une équirépartition des instances d'un père vers ses fils à nécessité la mise en place d'un algorithme spécifique qui offre une convergence rapide vers la solution recherchée.

Nous avons repris l'analogie proposée précédemment de manière à traiter le cas de l'héritage multiple. Nous avons également discuté de la prise en compte de l'héritage multiple à travers les mesures existantes. Nous avons montré que l'inégalité de Maguitman n'est plus respectée lorsque l'héritage multiple est autorisé. Notre proposition d'un cadre fédérateur pour la définition des mesures sémantiques ainsi adapté pour l'héritage multiple permet, comme nous l'avons évoqué en dernier lieu, de généraliser l'approche au calcul de la similarité entre deux sous-ensembles de concepts.

Troisième partie

Évaluation

7 Validation statistique

Sommaire

Résumé

La validation d'une mesure sémantique consiste à montrer qu'il s'agit de la mesure la plus adaptée à nos besoins. Il s'agit d'une tâche difficile que Budanitsky décompose en trois volets : l'analyse formelle, la comparaison avec le jugement humain et l'évaluation applicative. Nous introduisons ce chapitre en discutant de ces trois approches complémentaires. Nous poursuivons avec une présentation du réseau sémantique WordNet dont nous utilisons la structure hiérarchique pour nos expérimentations du fait de sa taille importante qui permet des études statistiques fiables. La suite de ce chapitre traite de l'intérêt de notre approche qui permet de se passer du corpus. Nous proposons une analyse statistique qui vise à évaluer la part d'information extraite du corpus et celle extraite de la structure hiérarchique. Nous complétons cette analyse par le biais de diverses comparaisons avec le jugement humain sur des jeux de tests (e.g. Miller&Charles) qui font référence dans la littérature. Ceci nous permet d'évaluer la pertinence de l'information extraite du corpus.

7.1 Introduction

Cette thèse présente une étude théorique aussi objective que possible menée dans un premier temps indépendamment de tout objectif applicatif précis. La pertinence de nos propositions peut toutefois être mise en évidence sur la base de comparaisons statistiques sur une hiérarchie de subsomption conséquente comme celle de WordNet [Fel98].

Nous traitons de la validation d'une mesure sémantique en discutant les trois approches complémentaires envisagées par Budanitsky [Bud99] :

- L'analyse formelle ;
- La comparaison avec le jugement humain ;
- L'évaluation applicative.

Ce chapitre traite également de l'évaluation de la part d'information extraite du corpus et de celle extraite de la structure hiérarchique. Il s'agit d'un point important qui a motivé nos travaux sur les approximations présentées aux chapitres 4 et 6. Pour cela, nous proposons plusieurs expérimentations utilisant WordNet qui de part sa taille nous permet de faire une étude statistique fiable.

7.2 La validation d'une mesure sémantique

Quelque soit le domaine de recherche dans lequel une mesure est définie, se pose la question de sa validation consistant à mettre en évidence l'adéquation de sa formalisation avec la signification recherchée. Pour évaluer une mesure, plusieurs approches complémentaires sont envisageables [Bud99] :

L'analyse formelle. On vise à étudier précisément leurs propriétés théoriques (métriques, ordinales, etc.) et leur distribution statistique.

Il s'agit d'étudier la mesure hors de tout contexte applicatif en isolant ses propriétés mathématiques et l'objet sur lequel elle porte. Cela permet de cerner l'information exploitée. En complément, on peut comparer plusieurs mesures afin de comprendre les relations qu'elles entretiennent et les corrélations qui en découlent.

La comparaison avec le jugement humain. Le principe est d'analyser la corrélation entre les valeurs de la mesure et les évaluations subjectives de sujets humains.

Les connaissances présentes dans l'ontologie ainsi que dans les autres sources d'information sont nécessairement incomplètes, imprécises et contiennent fort probablement des erreurs. On a finalement une modélisation partielle des connaissances de l'homme sur un domaine. Par conséquent, puisque les connaissances modélisées sont une vue partielle de ce qu'utilisent les individus qui formulent un jugement, il est souvent délicat de tirer des conclusions d'une petite variation de corrélation. Toutefois, la corrélation avec ces jugements humains reste un indicateur intéressant.

Certaines mesures [JC97, LBM03] proposent la pondération de certains éléments d'information à l'aide de coefficients (α, β, etc.). Ces contributions sont définies de manière à estimer « au mieux » le jugement humain. Dans ce cadre, le jugement humain, en plus d'être le référentiel de comparaison, devient une source d'information de la mesure. Finalement, on

se sert souvent du jeu de tests comme jeu d'apprentissage, ce qui nous semble abusif d'un point de vue méthodologique.

Cette approche nécessite également de disposer d'un échantillon statistiquement représentatif. La comparaison des mesures sémantiques sur le jeu de tests de Miller & Charles [MC91] qui contient 30 paires de concepts est très utilisée dans la littérature pour montrer l'apport des nouvelles mesures proposées. Il s'agit d'un sous-ensemble des 65 paires de Rubenstein & Goodenough[RG65]. Finkelstein & Gabrilovich [FGM+02] ont récemment proposé un jeu de tests plus conséquent contenant 350 paires dont celles de Rubenstein & Goodenough.

Les conclusions portent souvent sur des écarts de corrélations très faibles tandis que la corrélation linéaire de Pearson qui est utilisée est sensible au problème de représentation partielle des connaissances évoqué précédemment. Un indicateur moins ambitieux puisqu'il ne considère que le préordre induit par les mesures, mais plus fiable est la corrélation des rangs de Spearman. De plus, dans certaines applications, l'ordre des paires de concepts est parfois suffisant.

L'évaluation applicative. L'expérimentation est restreinte à un cadre applicatif bien identifié.

Lors de l'évaluation d'une mesure dans l'application visée il y a parfois beaucoup de paramètres qui entrent en jeu et rendent d'autant plus difficile l'analyse des résultats. Conjointement à une analyse théorique, l'analyse des résultats est plus aisée et permet d'affiner la description du comportement attendu de la mesure dans le contexte applicatif.

Ces trois approches sont complémentaires et participent à la connaissance des mesures et à la définition de la signification de la mesure nécessaire à l'application visée. Cette thèse est un support à l'analyse formelle des mesures sémantiques et nous ne cherchons pas à montrer la supériorité d'une mesure sur une autre dans une application précise. Cependant, nous donnons dans la section suivante quelques éléments qui mettent en évidence la pertinence de l'approche qui consiste à utiliser uniquement la structure hiérarchique lorsque le recours à un corpus n'est pas souhaité voire impossible. Pour cela, nous utilisons WordNet 2.0 qui étant un référentiel conséquent nous permet de réaliser des calculs statistiques fiables.

7.3 Qu'est-ce que WordNet ?

WordNet est un référentiel en ligne dont le développement est inspiré par des théories actuelles en psycholinguistique. Les noms anglais, les verbes et les adjectifs y sont organisés en ensembles de synonymes. Différentes relations lient ces ensembles de synonymes. WordNet vise finalement à modéliser les connaissances lexicales d'une personne dont la langue maternelle est l'anglais. Pour se faire, l'idée sous-jacente est de se rapprocher de l'organisation des connaissances lexicales de l'homme. C'est pourquoi cette initiative est basée sur des théories psycholinguistiques qui concernent l'organisation de la mémoire lexicale humaine [MBF+90].

Classiquement, l'organisation des informations lexicales se fait à l'aide de dictionnaires. Ceux-ci regroupent les mots qui ont une orthographe similaire tandis que les mots ayant une signification proche sont dispersés. Malheureusement, il n'y a pas d'alternative simple pour rechercher un mot sans y passer un peu de temps. L'outil informatique a permis d'apporter une réponse à ce problème. L'ordinateur est utilisé pour faire des recherches instantanées dans des dictionnaires informatisés comme il en existe sur la toile. Cependant, il est rapidement apparu qu'il était grossièrement réducteur d'utiliser des machines aussi puissantes comme de simples "tourneuses de pages rapides". Le problème est donc de trouver ce qu'on pourrait bien leur faire faire de plus. WordNet est une proposition qui tend à répondre à cette question. Finalement, WordNet s'inscrit dans le courant de la gestion des connaissances qui vise à proposer des solutions pour stocker des connaissances sur un support informatique afin de pourvoir ces machines de capacités de raisonnement.

Le Murray's Oxford English Dictionary a été créé selon des principes historiques et personne ne remet en doute la valeur de ce dictionnaire concernant la clarté des explications sur l'utilisation des mots. Cependant, en se focalisant sur des considérations historiques (diachroniques), les dictionnaires standards négligent les questions concernant l'organisation synchronique des connaissances lexicales. Ces dictionnaires sont issus d'une étude de l'évolution des termes plutôt que d'une étude des rapports entre termes coexistants d'un état de la langue. Cette lacune peut aujourd'hui être comblée. Le 20ème siècle a vu l'émergence de la psycholinguistique qui ouvre un champ de recherche interdisciplinaire concernant les bases cognitives des compétences linguistiques. Les psycholinguistes ont découvert de nombreuses propriétés du lexique mental qui peuvent être exploitées en lexicographie. En 1985 un groupe de psychologues et de linguistes à l'université de Princeton ont entrepris de développer une base de données lexicale selon les idées de Miller [Mil85]. L'idée initiale était de fournir une aide pour la recherche conceptuelle dans un dictionnaire. Pour cela la base de données serait utilisée conjointement avec un dictionnaire en ligne conventionnel. WordNet est le résultat de ces travaux.

Une différence fondamentale entre un dictionnaire classique et WordNet est que WordNet divise l'ensemble du lexique en 5 catégories : noms, verbes, adjectifs, adverbes et mots fonctionnels. En réalité, WordNet contient uniquement les noms, verbes, adjectifs et adverbes. L'ensemble relativement restreint des mots fonctionnels anglais est omis du fait que ceux-ci sont (selon certaines observations sur des patients aphasiques) probablement stockés séparément comme partie des composants syntaxiques du langage. C'est l'étude des associations de mots qui a mis en évidence le fait que les catégories syntaxiques diffèrent dans leur organisation subjective. Fillenbaum and Jones [FJ65] ont demandé à des sujets parlant l'anglais de donner le premier mot auquel ils pensaient en réponse à des mots tirés de différentes catégories syntaxiques. Le mot donné en réponse et celui proposé appartiennent majoritairement à la même catégorie : un nom entraîne un nom en réponse dans 79% des cas, un adjectif entraîne un adjectif en réponse dans 65% des cas, un verbe entraîne un verbe en réponse dans 43% des cas.

Puisque WordNet est censé être organisé selon les principes propres à la mémoire lexicale humaine, la décision d'organiser les noms en une hiérarchie

reflète un jugement psycholinguistique à propos du lexique mental. La plupart des psycholinguistes sont d'accord sur le fait que les noms anglais sont organisés hiérarchiquement dans la mémoire sémantique, mais sur le fait que les informations génériques soient stockées de façon redondante ou héritées est discutable [Smi78]. Collins et Quillian [CQ69] ont fait des expérimentations qui les ont amenés à conclure que les informations génériques n'étaient pas stockées de façon redondante mais retrouvées au besoin. En revanche, d'autres psycholinguistes ne sont pas d'accord avec ces conclusions. Dans WordNet, les noms sont organisés selon une hiérarchie.

La plupart des recherches ayant un intérêt en psycholexicologie utilise un sous-ensemble restreint du lexique anglais et souvent concentré uniquement sur les noms. Une des motivations pour le développement de WordNet est de fournir une extension complète de ce lexique. Par exemple, dans sa version 2.0, WordNet renferme 141690 paires mot-sens pour les noms, 24632 pour les verbes, 31015 pour les adjectifs et 5808 pour les adverbes.

Dans nos expérimentations, nous utilisons la hiérarhie des noms de WordNet qui nous permet par sa taille de tirer des conclusions probantes. Pour cela, nous nous appuyons sur l'application développée par Pedersen et al. [PPM04].

7.4 Intérêt de notre approche

L'objet premier de cette thèse est de proposer un cadre fédérateur pour la définition des mesures sémantiques comme nous l'avons vu au chapitre 5. Il serait illusoire et abusif de chercher à mettre en évidence l'intérêt de notre approche en choisissant une application spécifique. L'intérêt de notre proposition réside dans une approche théorique indépendante de tout objectif applicatif précis. Cela nous permet de fournir une analyse objective qui n'est pas biaisée par la recherche d'un objectif donné.

Nous allons toutefois donner quelques éléments de réflexions sur l'exploitation d'une hiérarchie en l'absence de corpus. Pour nos diverses expérimentations, nous avons repris l'application de Pedersen et al. [PPM04] à laquelle nous avons ajouté un certain nombre de modules développés en Perl. Cette application permet l'utilisation de plusieurs corpus dont un corpus conséquent, le « British National Corpus » que nous utilisons dans nos expérimentations. Diverses possibilités sont offertes pour la comptabilisation des occurrences, nous utilisons la méthode de comptage de Resnik[1] avec un lissage[2] de 1.

Nous allons dans un premier temps essayer d'évaluer la masse d'information qui est réellement extraite du corpus et celle qui vient de la structure hiérarchique. Dans un deuxième temps, nous évaluerons la pertinence de la masse d'information issue du corpus lorsqu'il s'agit d'approcher le jugement humain.

[1] Chaque concept associé à un mot reçoit une part équivalente de l'occurrence. Par exemple, si il y a deux sens d'un mot w, alors quand on observe w dans le corpus chaque concept associé à chaque sens voit son nombre d'occurrence augmenté de 0.5.

[2] Tous les concepts recoivent une occurrence au départ de manière à éviter que certains concepts n'ai aucune occurrence.

7.4.1 Apport du corpus

Pour mettre en évidence l'importance de la masse d'information extraite du corpus par rapport à celle qui peut être extraite de la structure hiérarchique, nous avons mené quatre expérimentations. Ces expérimentations visent à étudier la corrélation entre le contenu informationnel d'un concept obtenu avec l'approximation utilisant le corpus ($\widehat{\mathrm{P}}_c$) et diverses approximations basées uniquement sur la hiérarchie. Nous n'utilisons pas les jugements humains mais faisons uniquement des corrélations entre contenus informationnels utilisant différentes approximations. La racine est toujours considérée comme virtuelle avec $\psi(c_0) = 0$.

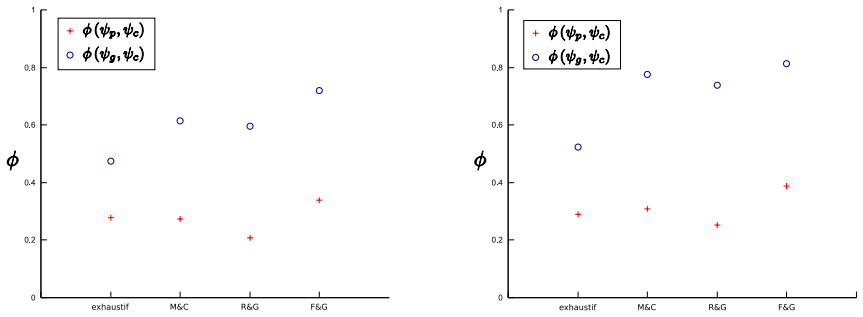
L'algorithme du calcul des fréquences d'occurrences proposé par Resnik induit une exploitation ascendante de la hiérarchie de subsomption proche de l'approximation $\widehat{\mathrm{P}}_g$ et de ses diverses variantes qui permettent de relaxer la propriété de complétude. Nous présentons donc les résultats obtenus avec chacune de ces variantes ainsi qu'avec l'approximation $\widehat{\mathrm{P}}_p$ qui met en évidence les différences significatives des approches descendantes avec les approches ascendantes.

Pour la première expérimentation, nous avons calculé le contenu informationnel de l'ensemble des 141690 concepts de WordNet avec les diverses approximations. Ensuite, nous avons restreint cet ensemble de concepts à ceux utilisés dans le jeu de tests de Miller & Charles [MC91] (cf. tableau 7.1).

car	automobile	3.92	crane	implement	1.66
gem	jewel	3.84	journey	car	1.16
journey	voyage	3.84	monk	oracle	1.10
boy	lad	3.76	cemetery	woodland	0.95
coast	shore	3.70	food	rooster	0.89
asylum	madhouse	3.61	coast	hill	0.87
magician	wizard	3.50	forest	graveyard	0.84
midday	noon	3.42	shore	woodland	0.63
furnace	stove	3.11	monk	slave	0.55
food	fruit	3.08	coast	forest	0.42
bird	cock	3.05	lad	wizard	0.42
bird	crane	2.97	chord	smile	0.13
tool	implement	2.95	glass	magician	0.11
brother	monk	2.82	noon	string	0.08
lad	brother	1.68	rooster	voyage	0.08

TAB. 7.1 – Jeu de tests de Miller et Charles

Nous avons ensuite fait de même avec ceux utilisés par Rubenstein & Goodenough [RG65] puis avec ceux de Finkelstein & Gabrilovich [FGM+02]. La figure 7.1 présente respectivement les comparaisons avec la corrélation des rangs de Spearman et la corrélation linéaire de Pearson. Les corrélations obtenues avec les variantes de $\widehat{\mathrm{P}}_g$ sont dans l'ensemble légèrement supérieures à celles obtenues avec la version initiale de $\widehat{\mathrm{P}}_g$ qui respecte la propriété de complétude, c'est pourquoi nous ne les faisons pas apparaitre sur les graphiques. La corrélation des rangs est systématiquement légèrement inférieure à la corrélation linéaire, mais cela n'influence pas les conclusions que l'on peut tirer de ces expérimentations.



(Corrélation des rangs) (Corrélation linéaire)

FIG. 7.1 – Corrélation entre les contenus informationnels obtenus avec et sans corpus

Remarquons tout d'abord que les divers jeux d'essais utilisent des échantillons de concepts qui ne sont pas si représentatifs que cela de l'ensemble des concepts. En effet, l'influence du corpus sur le contenu informationnel des concepts est plus important sur l'ensemble des concepts que dans chacun des jeux de tests. Cela a nécessairement un impact (de manière positive ou négative) sur les résultats des mesures utilisant le corpus. Selon ce point de vue, le jeu de tests de Finkelstein & Gabrilovich [FGM+02] est le plus dégradé des trois.

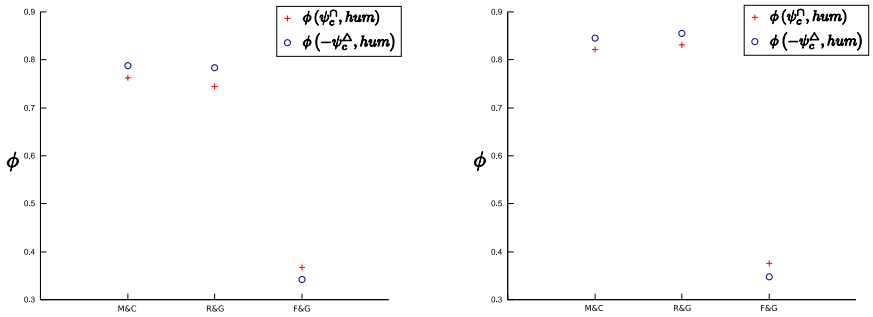
De manière prévisible, le contenu informationnel basé sur l'approximation $\widehat{\mathrm{P}}_p$ est le moins corrélé avec celui qui exploite le corpus. Les corrélations obtenues traduisent toutefois la relation évidente entre les méthodes ascendantes et descendantes dans le sens où la profondeur à tendance à être inversement proportionnelle à la hauteur.

La conclusion sans doute la plus importante que l'on peut tirer de ces expérimentations, c'est que les corrélations entre ψ_g et ψ_c montre que la masse d'information extraite du corpus en plus de celle inhérente à la structure de la hiérarchie est relativement restreinte. Ces résultats sont dépendants du corpus, mais surtout de la structure de WordNet. Sur une autre hiérarchie et un autre corpus, les résultats pourraient être assez différents à l'avantage de l'utilisation du corpus ou au contraire en montrant son inutilité. C'est pourquoi nous nous garderons de généraliser nos conclusions à l'ensemble des hiérarchies et corpus disponibles.

La masse d'information aussi modeste soit-elle d'un point de vue quantitatif peut être très pertinente et ainsi améliorer significativement la performance des mesures qui l'intègrent. Nous allons étudier la pertinence de cette masse d'information lorsque l'objectif est d'approcher le jugement humain.

7.4.2 Pertinence du corpus

Pour évaluer la pertinence de l'information extraite du corpus, nous regardons tout d'abord la contribution du contenu informationnel partagé $\psi_c^{\cap}(c_i, c_j)$ par deux concepts c_i et c_j ainsi que le contenu informationnel de ce qui les



(Corrélation des rangs) (Corrélation linéaire)

FIG. 7.2 – Contributions de $\psi_c^{\cap}$ et $\psi_c^{\triangle}$ pour approcher le jugement humain

distingue $\psi_c(c_i) + \psi_c(c_j) - 2\psi_c^{\cap}(c_i, c_j)$ (on parlera de contenu informationnel différentiel que l'on note $\psi_c^{\triangle}(c_i, c_j)$). Pour cela, nous avons calculé les corrélations avec le jugement humain sur les trois jeux de tests déjà évoqués.

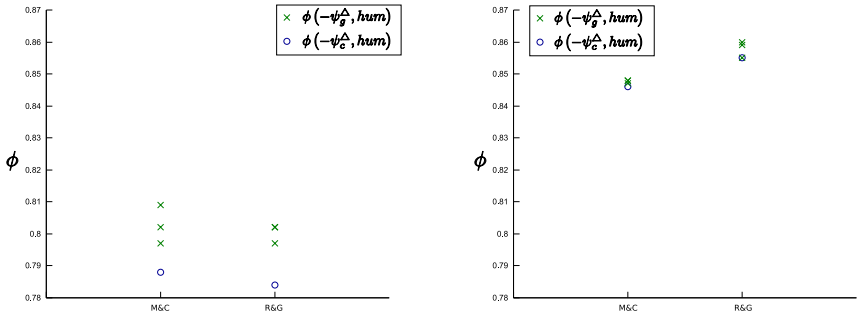
Les résultats présentés sur la figure 7.2 montrent que la contribution de $\psi_c^{\triangle}$ est plus importante que celle de $\psi_c^{\cap}$ pour les jeux de tests de Miller & Charles et Rubenstein & Goodenough tandis que la constatation inverse peut être faite pour le jeu de tests de Finkelstein & Gabrilovich. Cette tendance semble traduire une sensibilité différente des individus qui peut être dûe à la procédure d'évaluation mise en place lors de ces différents jeux de tests. Il s'agit peut-être également de l'influence de l'écart entre les connaissances modélisées et celles des individus qui ne sont pas les mêmes dans les diverses expériences. Cette fluctuation montre quoiqu'il arrive les limites de cette technique de validation.

En réitérant l'expérimentation précédente avec cette fois-ci l'approximation ψ_g, les conclusions sont les mêmes. De plus, si on compare les corrélations obtenues avec le corpus (grâce à ψ_c) et sans le corpus (grâce à ψ_g), on s'aperçoit que les corrélations sont très proches et globalement les écarts négligeables. Cela tend à montrer que l'information extraite du corpus est fortement bruitée.

Si on regarde de plus prêt, il est même intéressant de constater que si on considère la composante qui contribue le plus à approcher les jugements humains, l'écart de corrélation est toujours à l'avantage de l'approximation ψ_g et de ses variantes qui relaxent la complétude (cf. figures 7.3 et 7.4). Ces quelques expérimentations montrent que l'utilisation du corpus n'est pas toujours nécessaire voire judicieuse. Les diverses approximations que nous proposons au cours des chapitres 4 et 6 sont autant de possibilités qui permettent de s'abstraire du corpus.

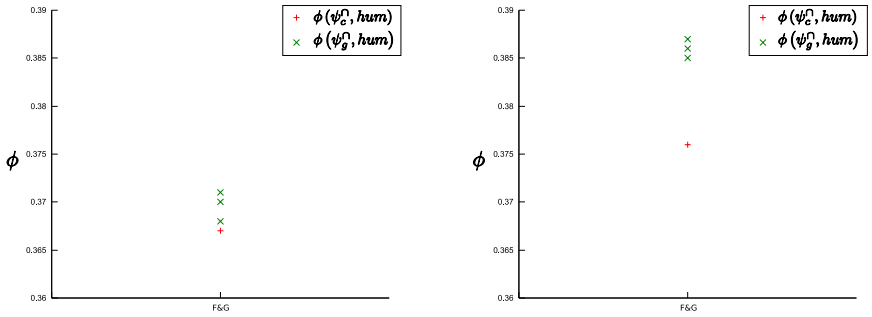
7.5 Conclusion

Dans ce chapitre, nous avons repris et discuté les trois approches de Budanitsky [Bud99] concernant la validation d'une mesure sémantique (l'analyse formelle, la comparaison avec le jugement humain et l'évaluation applicative).



(Corrélation des rangs) (Corrélation linéaire)

FIG. 7.3 – Comparaison des contributions de $\psi_c^{\triangle}$ et $\psi_g^{\triangle}$ pour approcher le jugement humain



(Corrélation des rangs) (Corrélation linéaire)

FIG. 7.4 – Comparaison des contributions de $\psi_c^{\cap}$ et $\psi_g^{\cap}$ pour approcher le jugement humain

Nous avons notamment mis en avant la complémentarité de ces trois approches et des difficultés rencontrées pour la comparaison avec le jugement humain.

Nous avons traité un aspect important de notre approche qui donne les clés pour exploiter une hiérarchie sans recourir à un corpus. Nous avons montré que dans le cas de WordNet, un référentiel de connaissances très utilisé, la part d'information extraite du corpus n'est pas très importante et surtout n'est pas très pertinente lorsqu'il s'agit d'approcher le jugement humain.

Etude des similarités sémantiques avec SymanticTab

8

Sommaire

Résumé

Le choix d'une mesure sémantique nécessite de comparer les mesures en les appliquant à la hiérarchie de subsomption utilisée dans l'application visée. La difficulté d'implémenter de nombreuses mesures contraint généralement à restreindre d'emblée les mesures potentielles. Nous proposons un plug-in pour l'éditeur d'ontologies Protégé2000 baptisé SymanticTab qui implémente l'approche soutenue dans cette thèse. Nous évoquons également une adaptation de ce plug-in pour un cadre applicatif spécifique lié à la base de connaissances UEML (*Unified Enterprise Modelling Language*). Nous proposons un cas d'utilisation de SymanticTab sur une hiérarchie de subsomption de taille restreinte et sans héritage multiple. Nous

détaillons également les points essentiels de l'implémentation de notre plug-in et de son adaptation pour UEML.

8.1 Introduction

Nous avons proposé dans les chapitres précédents un cadre pour définir des mesures de similarité sémantiques en choisissant : la forme de la mesure, l'approximation de la mesure de probabilité et le statut de la racine. Toutefois, l'impact de ce paramétrage pour une application donnée n'est pas toujours une tâche facile pour l'utilisateur.

Dans ce chapitre, nous présentons SymanticTab, un outil que nous avons développé pour définir diverses mesures (en faisant varier les paramètres) de manière à les appliquer à un domaine donné. La comparaison de leurs résultats aide l'utilisateur à comprendre leurs variations en fonction des valeurs de ces paramètres et à choisir celles qui s'adaptent au mieux à son application.

Nous traitons également de l'adaptation et de l'extension de SymanticTab pour la définition de mesures pour la comparaison de sous-ensembles de concepts, afin de comparer les constructs des langages de modélisation d'entreprise, dans le cadre de l'approche UEML dans le Réseau d'excellence Interop[1].

Enfin, nous exposons le choix de développer SymanticTab sous la forme d'un plug-in de Protégé2000 et nous présentons son architecture.

8.2 Fonctionnalités de SymanticTab

Nous avons développé l'outil SymanticTab pour aider l'utilisateur à définir rapidemment plusieurs mesures pour une hiérarchie donnée. Cet outil baptisé SymanticTab se présente sous la forme d'un plug-in pour Protégé2000 (cf. paragraphe 8.5.1). L'interface utilisateur du plug-in SymanticTab se décompose en trois parties (cf. figure 8.1) :

- la visualisation de la hiérarchie de subsomption à considérer ;
- la définition des similarités sémantiques ;
- l'analyse des similarités sur la hiérarchie de subsomption.

Protégé propose déjà la visualisation de la hiérarchie de concepts dans l'onglet *Classes*. Néanmoins, le fait de reprendre cette visualisation au niveau de l'interface de notre plug-in facilite l'analyse des similarités. L'utilisateur n'a plus à naviguer entre l'interface du plug-in et celle de Protégé.

Si l'on considère le sens de lecture de l'utilisateur (de gauche à droite et de haut en bas), la partie sur la définition des similarités sémantiques fait suite à la visualisation de la hiérarchie. Enfin, la partie destinée à l'analyse des similarités définies par l'utilisateur couvre le reste de l'interface.

[1] www.interop-noe.org

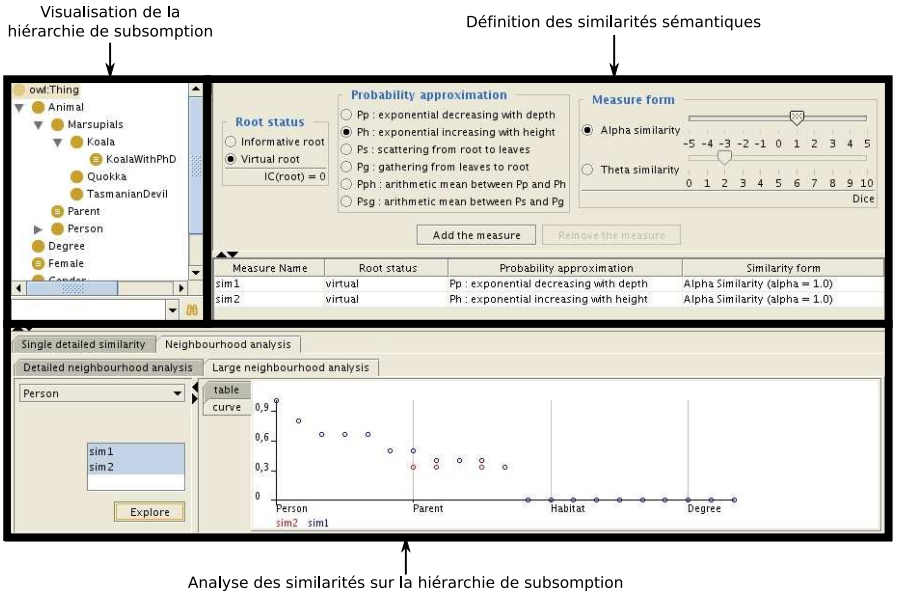


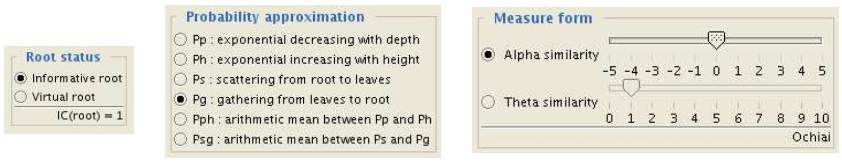
FIG. 8.1 – Interface globale du plug-in SymanticTab

Visualisation de la hiérarchie

Le panneau destiné à la visualisation de la hiérarchie fournit une représentation arborescente de celle-ci. Le cas de l'héritage multiple est traité à l'aide d'une recopie de la sous-arborescence concernée au niveau de chacun des parents. Cette interface de visualisation fournit également une zone de saisie pour rechercher rapidement un concept dans la hiérarchie.

Définition des similarités

Le panneau dédié à la définition des mesures de similarité distingue trois aspects pour le paramétrage des similarités. Il s'agit tout d'abord de définir le statut de la racine dont dépend son contenu informationnel. Comme le montre la figure 8.2.A, le contenu informationnel correspondant est affiché en bas à droite (IC(root)=1).



(A) (B) (C)

FIG. 8.2 – Paramétrage pour la définition d'une mesure

Le choix de l'une des diverses approximations que nous avons présentées au

cours des chapitres 4 et 6 intervient également dans le paramétrage proposé par l'interface (cf. figure 8.2.B).

Enfin, il est possible de définir la forme de la mesure en choisissant une mesure de type α ou θ. Sur l'exemple de la figure 8.2.C, une mesure de type α a été choisie avec $\alpha = 0$. En bas à droite s'affiche lorsque cela est possible le nom de la mesure à laquelle cela correspond (e.g. Jaccard, Dice, Ochiaï).

Lorsque l'utilisateur a terminé le paramétrage d'une mesure, un bouton permet d'ajouter celle-ci à la liste des mesures existantes (figure 8.3). Lorsque l'une des mesures de la liste est sélectionnée, le bouton de suppression devient accessible (cf. figure 8.4). La hiérarchie de subsomption et la liste de similarités ainsi définies sont utilisées par le troisième panneau pour l'analyse.

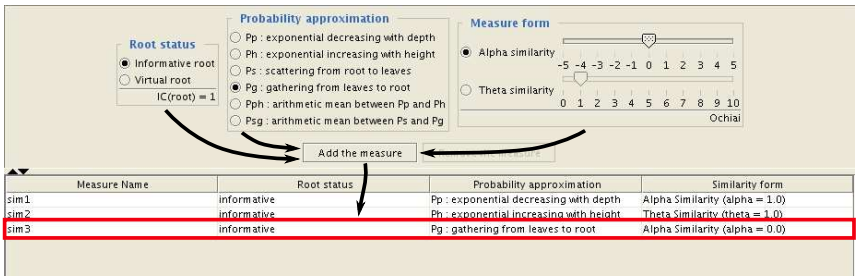


FIG. 8.3 – Définition d'une nouvelle similarité

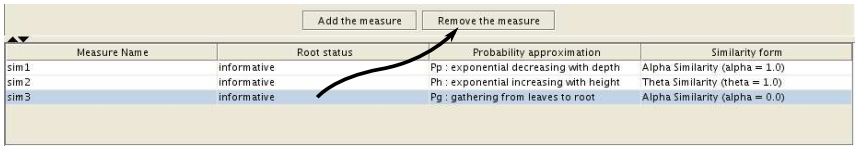


FIG. 8.4 – Suppression d'une similarité

Analyse des similarités

L'analyse des similarités peut se faire selon trois niveaux de détail :
- le calcul détaillé d'une similarité entre deux concepts (cf. figure 8.5)
- l'exploration du voisinage d'un concept selon une mesure de similarité (cf. figure 8.6)
- l'exploration du voisinage d'un concept selon plusieurs mesures de similarité (cf. figure 8.7)

Lors du calcul d'une similarité entre deux concepts comme d'ailleurs dans l'exploration du voisinage d'un concept, nous avons détaillé le contenu informationnel des concepts considérés (IC(Concept1) et IC(Concept2)) ainsi que leur contenu informationnel partagé (SIC(Concept1,Concept2)).

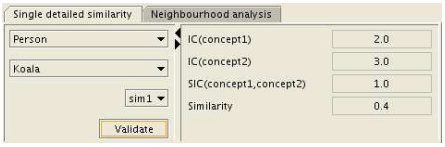


FIG. 8.5 – Calcul d'une similarité entre deux concepts

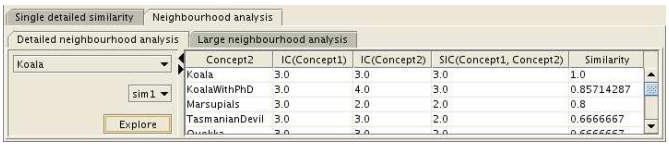


FIG. 8.6 – Calcul de la similarité entre un concept et tous les autres concepts à l'aide d'une seule mesure sémantique

L'exploration du voisinage avec plusieurs mesures donne des tableaux plus difficiles à exploiter. Nous proposons donc un second mode de visualisation sous forme de courbes 2D avec un tri suivant l'une des mesures. Cela permet de cibler rapidement les différences d'évaluation d'une mesure à l'autre.

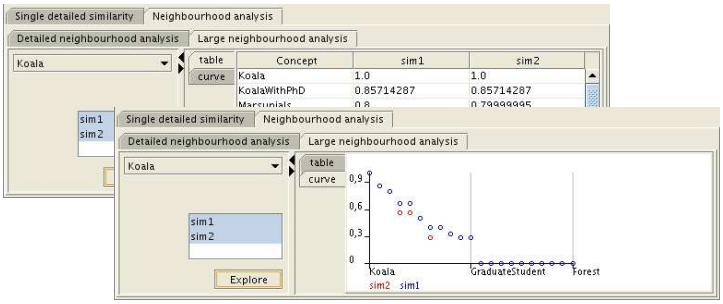


FIG. 8.7 – Calcul de la similarité entre un concept et tous les autres concepts avec plusieurs mesures sémantiques

Dans la suite, nous présentons l'extension de SymanticTab pour la comparaison des constructs des langages de modélisation d'entreprise, dans le cadre de l'approche UEML.

8.3 Adaptation pour UEML

Pour jouer un rôle important ou au moins survivre dans un monde économique en évolution permanente, les entreprises doivent avoir une vision claire de leur propre structure. La modélisation d'entreprise (*Enterprise Modelling*) est l'ensemble des activités et processus utilisés pour développer les diverses parties d'un modèle d'entreprise [PD02].

Les langages de modélisation d'entreprise (*Enterprise Modelling Languages*) permettent le développement de tels modèles d'entreprise. Un langage de modélisation d'entreprise définit les constructs génériques du modèle pour une modélisation d'entreprise adaptée aux besoins des gens qui créent et utilisent le modèle d'entreprise. Selon [Ver02], le nombre conséquent de langages de modélisation d'entreprise existants crée une situation difficile pour les utilisateurs désirant utiliser la modélisation d'entreprise qui peut être résumée ainsi :

- il y a trop de langages de modélisation d'entreprise pour les apprendre et les comprendre tous ainsi que trop d'outils avec des interfaces complètement différentes pour les maîtriser tous ;
- il y a un vocabulaire et des paradigmes de modélisation instables (le même concept peut avoir des noms différents et être modélisé différemment, tandis qu'un même terme peut faire référence à des choses différentes) ;
- il y a des incompatibilités entre les outils de modélisation d'entreprise sur le marché qui ne sont pas inter-opérables et peuvent difficilement échanger leur modèles ;
- il n'y a pas ou peu de fondements formels pour la modélisation d'entreprise.

Le réseau d'excellence INTEROP-NoE[2] (*Interoperability Research for Networked Enterprise Applications and Software*), est un projet qui s'est déroulé sur 42 mois (Novembre 2003 - Avril 2007) et coordonné par l'université de Bordeaux 1 avec 47 partenaires et plus de 300 chercheurs. C'est dans le cadre de ce projet qu'a émergé UEML (*Unified Enterprise Modelling Language*) qui cherche à faire face au problème de la multiplicité des langages de modélisation d'entreprise [BOAD04] [ABH+08].

8.3.1 Présentation d'UEML

L'objectif de UEML est de supporter l'utilisation intégrée des modèles d'entreprises définis dans des langages différents. UEML est conçue comme un mécanisme pour interconnecter des langages différents et leurs modèles. UEML comprend [OB06] :

- un méta-méta modèle pour organiser la description des différents aspects d'un construct ;
- une ontologie permettant de décrire la sémantique des constructs ;
- un cadre pour définir et évaluer la qualité des langages de modélisation d'entreprise pour aider à sélectionner les langages à décrire ;
- une approche d'analyse des correspondances pour déterminer les correspondances sémantiques entre les constructs ;
- un ensemble d'outils pour aider à l'utilisation des descriptions des langages considérés.

UEML propose une nouvelle approche pour décrire les langages de modélisation, leurs types de diagrammes et leurs constructs. L'objectif de cette approche est d'incorporer les langages existants et leurs constructs dans ce qui est nommé « web de langages », qui se définit comme un ensemble de langages sélectionnés par rapport à leur qualité d'une façon standardisée, intégrante et évolutive. Lors de la description d'un langage en informatique, on distingue souvent trois concepts : sa syntaxe concrète, sa syntaxe abstraite et sa sémantique. Pour les

[2] http ://www.interop-noe.org

langages de modélisation, ces trois concepts sont renommés respectivement : présentation (visualisation d'un construct), représentation (le méta-modèle de sa structure et ses relations) et alignement de la représentation (qui décrit comment les éléments de la représentation sont alignés avec les concepts de l'ontologie).

L'ontologie UEML a été développée à partir de l'ontologie de Bunge [Bun79] et le modèle BWW [WW93] puis enrichie avec d'autres concepts pour mieux couvrir la sémantique des constructs incorporés dans le web de langages de UEML. Elle a été conçue et structurée selon quatre concepts centraux : Classe, Propriété, Etat et Transformation. Chaque concept est spécialisé en d'autres concepts pour former une hiérarchie. Ces quatre hiérarchies sont liées par des relations entre leurs concepts. Les types de ces relations sont bien définis : une classe peut avoir une ou plusieurs propriétés, un état est défini par des propriétés et une transformation est définie par des états avant et après. Actuellement, les hiérarchies des classes (35 concepts) et propriétés (56 concepts) sont assez développées pour les exploiter, ce qui n'est pas le cas des deux autres hiérarchies (moins de 9 concepts par hiérarchie). La figure 8.8 représente la hiérarchie des classes de l'ontologie UEML après son enrichissement avec l'incorporation des langages suivants utilisés en modélisation d'entreprise : RdPC, GRL, KAOS, ISO/DIS 19440, UML 2.0, BPMN et IDEF3.

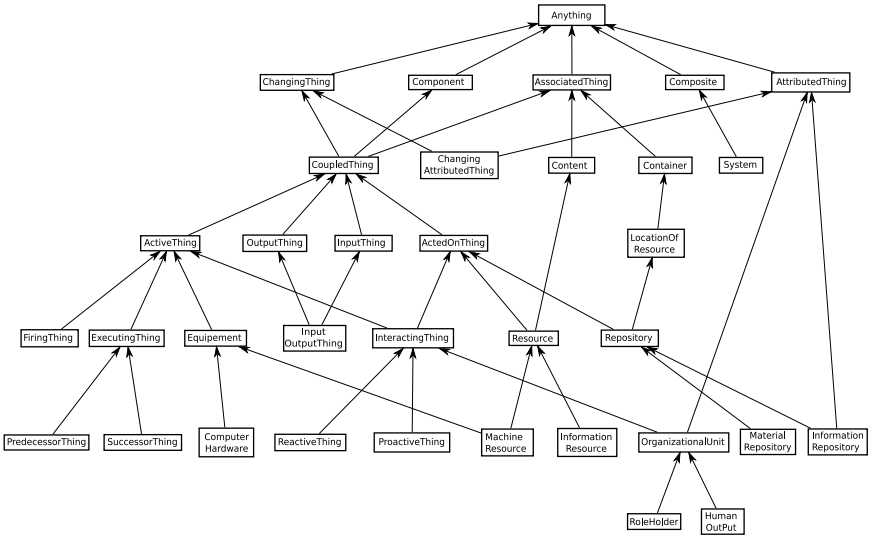


Fig. 8.8 – Hiérarchie des classes de UEML

Un construct d'un langage est donc intégré dans ce web de langages par la définition de sa présentation, sa représentation et l'alignement de cette dernière avec l'ontologie UEML. En comparant les alignements des représentations des constructs, on peut identifier les relations sémantiques entre ces constructs. Tous les constructs incorporés dans ce web de langages sont par conséquent inter-reliés au niveau le plus détaillé via l'ontologie UEML.

8.3.2 Mesures sémantiques entre deux constructs

La description d'un construct est définie par un graphe qui comprend des concepts des quatre hiérarchies de UEML (Classe, Propriété, Etat et Transformation). Pour évaluer une liaison entre deux constructs, nous les considérons comme des sous-ensembles de concepts de ces quatre hiérarchies. Lors du développement du plug-in SymanticTab, nous nous sommes restreints à la définition de similarités entre concepts. Nous avons adapté ce plug-in pour permettre l'analyse de mesures sémantiques (pas seulement des similarités) entre deux constructs (deux sous-ensembles de concepts) d'UEML. Une mesure entre deux constructs dans chacune des quatre hiérarchies donne quatre points de vue différents.

Pour tenir compte de ces quatre aspects à la fois, nous avons regroupé les quatre hiérarchies sous une même racine. On note respectivement $\mathcal{C}_{c_i}$, $\mathcal{C}_{p_i}$, $\mathcal{C}_{e_i}$ et $\mathcal{C}_{t_i}$ l'ensemble des concepts de la hiérarchie des classes, des propriétés, des états et des transformations liés au construct X_i. Dans ce cadre, le contenu informationnel global d'un construct $\psi^{\cup}(X_i)$ est équivalent à la somme des contenus informationnels globaux de chaque sous-ensemble de concepts :

$$\begin{aligned}\psi^{\cup}(X_i) &= \psi^{\cup}(\mathcal{C}_{c_i} \cup \mathcal{C}_{p_i} \cup \mathcal{C}_{e_i} \cup \mathcal{C}_{t_i}) \\ &= \psi^{\cup}(\mathcal{C}_{c_i}) + \psi^{\cup}(\mathcal{C}_{p_i}) + \psi^{\cup}(\mathcal{C}_{e_i}) + \psi^{\cup}(\mathcal{C}_{t_i})\end{aligned}$$

Le contenu informationnel partagé $\psi^{\cap}(X_i, X_j)$ par deux constructs X_i et X_j se résume à une somme de contenus informationnels partagés :

$$\begin{aligned}\psi^{\cap}(X_i, X_j) &= \psi^{\cap}(\mathcal{C}_{c_i} \cup \mathcal{C}_{p_i} \cup \mathcal{C}_{e_i} \cup \mathcal{C}_{t_i}, \mathcal{C}_{c_j} \cup \mathcal{C}_{p_j} \cup \mathcal{C}_{e_j} \cup \mathcal{C}_{t_j}) \\ &= \psi^{\cap}(\mathcal{C}_{c_i}, \mathcal{C}_{c_j}) + \psi^{\cap}(\mathcal{C}_{p_i}, \mathcal{C}_{p_j}) + \psi^{\cap}(\mathcal{C}_{e_i}, \mathcal{C}_{e_j}) + \psi^{\cap}(\mathcal{C}_{t_i}, \mathcal{C}_{t_j})\end{aligned}$$

La différence de taille entre les quatre hiérarchies entraîne un biais qui donne plus d'importance aux propriétés qu'aux classes tandis que les états et transformations n'ont que très peu d'impact sur la similarité. Il est donc nécessaire de normaliser chaque contenu informationnel à l'aide du contenu informationnel maximal correspondant. On note respectivement $\mathcal{C}_c$, $\mathcal{C}_p$, $\mathcal{C}_e$ et $\mathcal{C}_t$ l'ensemble des concepts de la hiérarchie des classes, des propriétés, des états et des transformations. Les facteurs d'échelle ϕ_c, ϕ_p, ϕ_e et ϕ_t associés respectivement aux classes, propriétés, états et transformations sont définis comme l'inverse des contenus informationnels maximaux correspondants :

$$\begin{aligned}\phi_c &= \frac{1}{\psi^{\cup}(\mathcal{C}_c)} \\ \phi_p &= \frac{1}{\psi^{\cup}(\mathcal{C}_p)} \\ \phi_e &= \frac{1}{\psi^{\cup}(\mathcal{C}_e)} \\ \phi_t &= \frac{1}{\psi^{\cup}(\mathcal{C}_t)}\end{aligned}$$

L'utilisateur détermine quatre coefficients w_c, w_p, w_e et w_t qui donnent l'influence des contenus informationnels issus de chacune des hiérarchies initiales. En définitive ces coefficients pondèrent les contenus informationnels de la même manière que les facteurs d'échelle.

8.4 Un cas d'utilisation

Nous proposons un cas d'utilisation du plug-in SymanticTab sur un arbre de subsomption contenant une vingtaine de concepts. Il s'agit du fichier *koala.owl* au format OWL écrit par Holger Knublauch[3] qui a participé au développement de Protégé. Nous avons chargé cette ontologie sous Protégé, ce qui nous permet d'en visualiser l'arbre de subsomption comme le montre la figure 8.9.

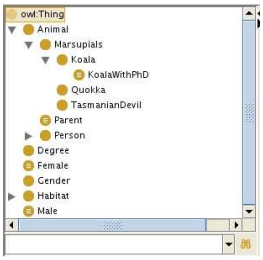


FIG. 8.9 – Visualisation de l'arbre de subsomption de koala.owl

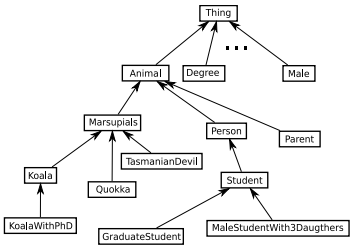


FIG. 8.10 – Représentation d'une partie de l'arbre de subsomption de koala.owl

Nous allons maintenant analyser successivement l'impact du statut de la racine, de l'approximation choisie et de la forme de la mesure à l'aide du plug-in Symantictab. Cette analyse porte sur une partie de la hiérarchie illustrée par la figure 8.10.

8.4.1 Statut de la racine

Nous avons tout d'abord défini deux similarités nommées sim1 et sim2 (cf. figure 8.11) ayant la même forme (similarité $\sigma_{\alpha=1}$) et utilisant la même approximation ($\widehat{\mathrm{P}}_p$). Tandis que sim1 considère la racine *Thing* comme informative ($\psi\,(Thing) = 1$), sim2 la considère comme virtuelle ($\psi\,(Thing) = 0$).

Measure Name	Root status	Probability approximation	Similarity form
sim1	informative	Pp : exponential decreasing with depth	Alpha Similarity (alpha = 1.0)
sim2	virtual	Pp : exponential decreasing with depth	Alpha Similarity (alpha = 1.0)

FIG. 8.11 – Définition de deux similarités qui diffèrent au regard de la prise en compte du statut de la racine

Dans la partie inférieure de l'interface destinée à l'analyse des mesures de similarités sur la hiérarchie, nous avons lancé le calcul des deux mesures préalablement définies entre le concept *Animal* et les autres concepts de l'arbre de subsomption. Nous obtenons les courbes telles qu'elles apparaissent sur la figure 8.12. Nous y avons mis en évidence à l'aide de flèches l'annulation de la similarité entre le concept *Animal* et d'autres concepts comme *Male* ou *Degree*. En effet, *Animal* n'ayant en commun avec ces concepts que la racine *Thing*, lorsque celle-ci est considérée comme virtuelle, leur similarité s'annule.

[3]http ://www.knublauch.com/

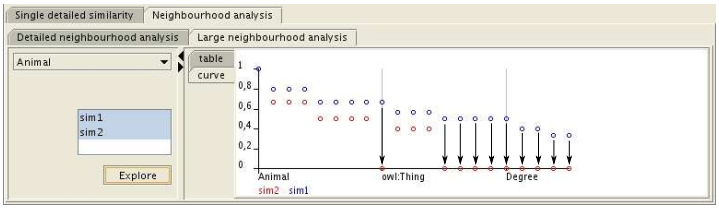


FIG. 8.12 – Courbes des similarités entre le concept *Animal* et les autres concepts selon les deux similarités définies

C'est au regard du comportement attendu que l'utilisateur pourra alors choisir dorénavant de considérer la racine comme informative ou virtuelle (Nous faisons ici arbitrairement le choix d'une racine virtuelle). Nous poursuivons cette analyse en présentant deux comparaisons de mesures qui peuvent guider le choix de l'approximation.

8.4.2 Approximation

Nous redéfinissons les mesures sim1 et sim2 (cf. figure 8.13) de manière à ce qu'elles considèrent la racine comme virtuelle et aient la même forme (similarité $\sigma_{\alpha=1}$). Tandis que sim1 repose sur l'approximation $\widehat{\mathrm{P}}_p$ relevant de l'approche descendante, sim2 repose sur l'approximation $\widehat{\mathrm{P}}_h$ relevant de l'approche ascendante.

Measure Name	Root status	Probability approximation	Similarity form
sim1	virtual	Pp : exponential decreasing with depth	Alpha Similarity (alpha = 1.0)
sim2	virtual	Ph : exponential increasing with height	Alpha Similarity (alpha = 1.0)

FIG. 8.13 – Définition de sim1 qui repose sur l'approximation $\widehat{\mathrm{P}}_p$ et sim2 qui repose sur l'approximation $\widehat{\mathrm{P}}_h$

Nous avons lancé le calcul des mesures sim1 et sim2 entre le concept *Person* et les autres concepts de l'arbre de subsomption. Les courbes obtenues (cf. figure 8.14) montrent un écart entre les valeurs de ces deux similarités pour les concepts *Quokka* et *TasmanianDevil* mais surtout pour le concept *Parent*.

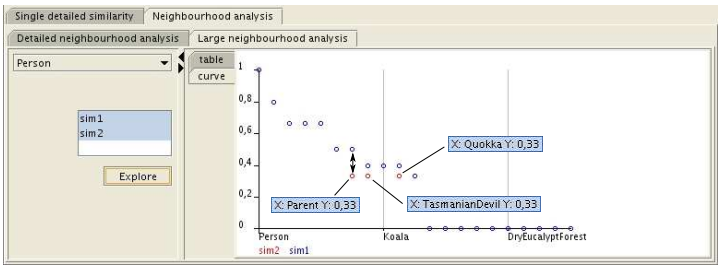


FIG. 8.14 – Courbes des similarités entre le concept *Person* et les autres concepts selon les deux mesures définies

Cet écart est dû au fait que ces trois concepts sont des feuilles donc considérés comme de spécificité maximale du point de vue de l'approche ascendante. En revanche l'approche descendante basée sur la profondeur leur donne une spécificité moins importante parce qu'ils ne sont pas de profondeur maximale. Pour mettre cela en évidence, nous pouvons consulter le détail du contenu informationnel de chaque concept et du contenu informationnel partagé par le biais de l'interface de notre plug-in comme le montre la figure 8.15. En effet, tandis que les deux similarités attribuent la même valeur de spécificité aux concepts *Person* et *Animal* (subsumant commun le plus spécifique), cela n'est pas le cas pour les trois concepts *Quokka*, *TasmanianDevil* et *Parent*. Considéré comme plus spécifique par sim2, ces trois concepts ont plus de différence avec le concept *Person*, ce qui explique une similarité plus faible.

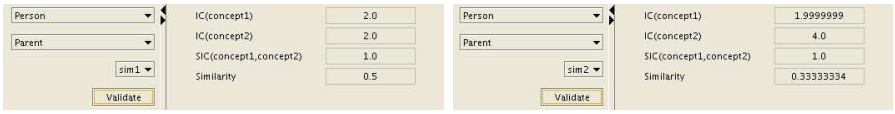


FIG. 8.15 – Détail du contenu informationnel des concepts *Person* et *Parent* nécessaire au calcul de sim1 et sim2

Après avoir confronté les approches ascendante et descendante, nous nous attardons sur la prise en considération ou non du nombre de fils de chaque concept. Pour cela, nous redéfinissons uniquement la mesure sim2 de manière à ce qu'elle repose désormais sur l'approximation $\widehat{\mathrm{P}}_s$ (cf. figure 8.16).

Measure Name	Root status	Probability approximation	Similarity form
sim1	virtual	Pp : exponential decreasing with depth	Alpha Similarity (alpha = 1.0)
sim2	virtual	Ps : scattering from root to leaves	Alpha Similarity (alpha = 1.0)

FIG. 8.16 – Définition de sim1 qui repose sur l'approximation $\widehat{\mathrm{P}}_p$ et sim2 qui repose sur l'approximation $\widehat{\mathrm{P}}_s$

Les courbes obtenues (cf. figure 8.17) montrent que la prise en compte du nombre de fils peut se révéler problématique. Le concept *koala* est subsumé uniquement par le concept *koalaWithPhD*. Le fait de prendre en compte le nombre de fils revient à considérer que puisqu'il n'y a pas de koala sans doctorat, c'est que tous les koalas ont un doctorat et que donc ces deux concepts sont équivalents. Bien que l'interface ne le permette pas, nous pourrions relaxer la contrainte de complétude. Nous choisirons pour la suite de ne pas tenir compte du nombre de fils.

8.4.3 Forme de la mesure

Nous définissons maintenant trois similarités de type α qui considèrent la racine comme virtuelle ; elles se basent sur l'approximation $\widehat{\mathrm{P}}_p$ et prennent respectivement 1,-4 et 5 comme valeur de α (cf. figure 8.18).

Nous avons lancé le calcul des trois similarités entre le concept *koala* et les autres concepts. Les courbes correspondantes (cf. figure 8.19) montre l'impact

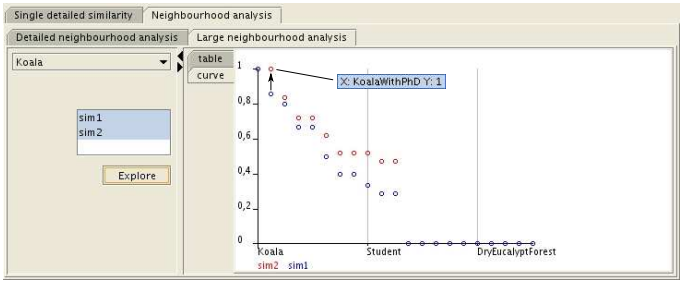


FIG. 8.17 – Courbes des similarités entre le concept *Koala* et les autres concepts selon les deux mesures définies

Measure Name	Root status	Probability approximation	Similarity form
sim1	virtual	Pp : exponential decreasing with depth	Alpha Similarity (alpha = 1.0)
sim2	virtual	Pp : exponential decreasing with depth	Alpha Similarity (alpha = -4.0)
sim3	virtual	Pp : exponential decreasing with depth	Alpha Similarity (alpha = 5.0)

FIG. 8.18 – Définition des similarités sim1, sim2 et sim3 de type α

de la différence de spécificité qui fait croître la similarité lorsque α diminue et qui fait décroître celle-ci lorsque α augmente.

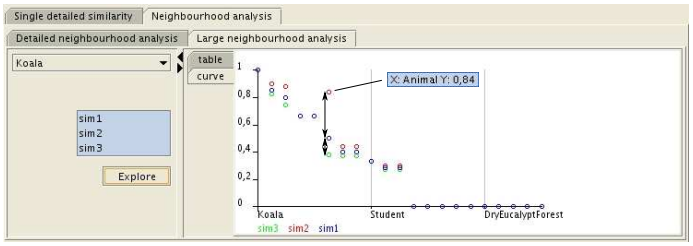


FIG. 8.19 – Courbes des similarités entre le concept *Koala* et les autres concepts selon les trois mesures définies

8.5 Implémentation

8.5.1 Protégé2000 : un éditeur d'ontologies

Parmi les nombreux outils d'édition d'ontologies existants, Protégé [NFM00] se démarque par la philosophie qui guide son développement. Il s'agit d'un système extensible destiné à être enrichi par ses utilisateurs grâce à l'ajout de nouvelles fonctionnalités sous forme de plug-ins.

Cet outil étant compatible avec tous les standards du web sémantique et en constante évolution, le développement d'un plug-in destiné à l'analyse et la comparaison de similarités sémantiques se révèle d'un grand intérêt. Notre plug-in est un support à l'analyse du comportement de diverses similarités sémantiques sur une hiérarchie de subsomption réelle. Il s'agit en définitive d'un

outil permettant à un utilisateur de choisir la mesure qui répond le mieux à sa problématique.

Le projet protégé2000 est issu d'une longue évolution dont l'origine remonte au système à base de connaissances de Mark Musen développé en 1987 [NFM00]. Il s'agit d'un outil d'édition d'ontologies développé au Stanford Medical Informatics de l'Université de Stanford. Protégé2000 a évolué pour devenir un système extensible pour le développement de systèmes à base de connaissances. La version actuelle de Protégé2000 (développée en java) peut être utilisée sur diverses platformes et permet une personnalisation de son interface. Elle reprend le modèle OKBC (*Open Knowledge Base Connectivity*) et utilise les formats de stockage standards comme les bases de données relationnelles, XML, RDF(S), OWL. Elle est utilisée par une communauté importante composée de milliers d'utilisateurs dont des groupes de recherche.

Le modèle de connaissances de Protégé2000 est issu du modèle des frames et contient des classes (concepts du domaine), des slots (propriétés des concepts), des facets (valeurs des propriétés), des axiomes (contrainte additionnelles) et des instances des classes [NFM00]. Protégé2000 se veut compatible avec le modèle OKBC [CFF+] mais en diffère sur certains points (e.g. le fait qu'une frame peut être une instance de plusieurs classes en OKBC et non sous Protégé) résumés dans [NFM00].

La figure 8.20 montre une ontologie éditée avec Protégé2000. Une partie de la hiérarchie de subsomption est visible en partie gauche de cette figure. Cet outil se distingue principalement des autres outils existants par son interface modulaire notamment grâce à son organisation sous la forme d'onglets.

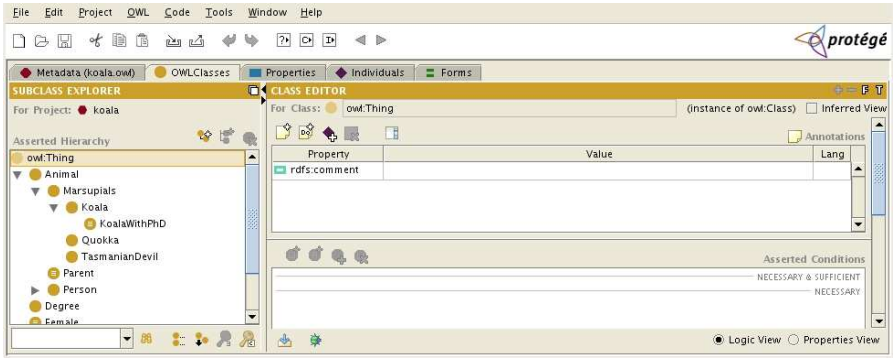


FIG. 8.20 – Interface de l'outil Protégé

8.5.2 Système de plug-ins

L'architecture de Protégé2000 part du principe que les utilisateurs veulent améliorer, personnaliser et tailler sur mesure le comportement du système par le biais de diverses extensions encore appelées « plug-ins ». Ces plug-ins sont des morceaux de code modulaires qui ajoutent de nouvelles fonctionnalités à Protégé2000. Il existe différents types de plug-ins dont les plug-ins de type *tab*

widget qui permettent l'ajout de nouveaux onglets à l'interface pour donner accès à de nouvelles fonctionnalités.

Il existe déjà de nombreux plug-ins qui ont grandement contribué à populariser cet outil et qui sont répertoriés sur le site officiel de Protégé[4]. Une API[5] Java permet de manipuler la base de connaissances de Protégé. Un objet de la classe *KnowledgeBase* accessible depuis le plug-in permet par exemple de retrouver la racine de la hiérarchie des classes. Le développement d'un plug-in de type *tabwidget* repose sur la classe abstraite *AbstractTabWidget*. L'exemple de code ci-dessous définit un plug-in nommé NouvelOnglet et qui permet d'étendre l'interface de Protégé comme le montre la figure 8.21.

```
import edu.stanford.smi.protege.widget.AbstractTabWidget;
import edu.stanford.smi.protege.model.KnowledgeBase;
import javax.swing.JButton;
import java.awt.FlowLayout;

public class NouvelOnglet extends AbstractTabWidget
{
    JButton monBouton;

    public void initialize()
    {
      // L'onglet est étiqueté
      this.setLabel("Nom du nouvel onglet");
      this.setLayout(new FlowLayout());
      // On dispose de la base de connaissances
      KnowledgeBase kb = this.getKnowledgeBase();
      // On utilise le nom de la racine comme texte du bouton
      monBouton = new JButton(kb.getRootCls().getName());
      this.add(monBouton);
    }
}
```

FIG. 8.21 – Extension de l'interface de Protégé à l'aide d'un nouvel onglet

Notre souhait étant de proposer une interface pour l'étude des similarités

[4]http ://protege.stanford.edu/
[5]Application Programming Interface

sémantiques sur une hiérarchie de subsomption, nous avons développé notre outil sous forme d'un plug-in de type *tabwidget* qui bénéficie de la puissance du logiciel Protégé. Pour utiliser le plug-in sur sa hiérarchie, l'utilisateur doit tout d'abord charger celle-ci dans Protégé. Ensuite, le menu Project->Configure lui permet de sélectionner le plug-in SymanticTab.

8.5.3 Réutilisation de composants Java

Le développement de l'interface a nécessité l'utilisation du *J2SE Software Development Kit (SDK)*. Nous avons notamment utilisé de nombreux widgets[6] de la librairie *Swing* : `JLabel`, `JButton`, `JTextField`, `JScrollPane`, `JSlider`, `JTable`, `JTabbedPane`, `JComboBox`, `JPanel`, `JSplitPane`, `JRadioButton`, `JSeparator`.

Pour l'affichage des courbes 2D, nous avons utilisé la librairie *JChart2D*[7] externe au SDK et distribuée sous licence GNU LGPL[8]. Elle est centrée autour du composant `Chart2D` qui étend la classe `JComponent` de la librairie *Swing* et s'intègre donc parfaitement à notre interface.

Nous avons effectué un découpage logique de notre interface de manière à obtenir un ensemble de classes cohérentes et chacune donnant lieu à un code de taille raisonnable. Nous avons réutilisé la vue `SelectClsesPanel` de l'API Protégé qui permet la visualisation de la hiérarchie. La figure 8.22 reprend l'arbre des widgets du plug-in SymanticTab qui fait apparaître les sous-parties `RootStatusView`, `ProbabilityApproximationView`, `MeasureFormView`, `SimilarityListView`, `SingleDetailedSimilarityView`, `DetailedNeighbourhoodAnalysisView`, `LargeNeighbourhoodAnalysisView`. Chacune de ces sous-parties donne lieu à une sous-arboresence de widgets de taille raisonnable.

La librairie *AWT* fournit plusieurs gestionnaires de mise en page (*Layout*) applicables notamment à un objet de la classe `JPanel`. Le gestionnaire `GridBagLayout` offre une grande précision dans le positionnement des widgets qui limite la profondeur de l'arbre des widgets. Son utilisation est fastidieuse et aboutit à une structure linéaire avec un code difficile à maintenir. Nous avons donc fait le choix d'utiliser une combinaison de gestionnaires plus simple de mise en oeuvre tel que `FlowLayout` ou `BorderLayout`. Nous avons également utilisé le widget `JSplitPane` qui permet de découper un panneau verticalement ou horizontalement en deux parties qui peuvent être masquées sur un simple clic de souris. La succession des gestionnaires utilisés fixe un minimum de contraintes et participent avec les widgets `JSplitPane` à la souplesse de l'interface.

8.5.4 Architecture MVC

La triade de classes Modèle/Vue/Controleur (MVC) a été proposée pour la construction d'interfaces en Smalltalk-80 [KP88]. MVC définit trois types d'objets [GHJV95] :

[6] Un widget est un composant d'interface graphique
[7] http ://jchart2d.sourceforge.net/
[8] Lesser General Public License (http ://www.gnu.org/copyleft/lesser.txt)

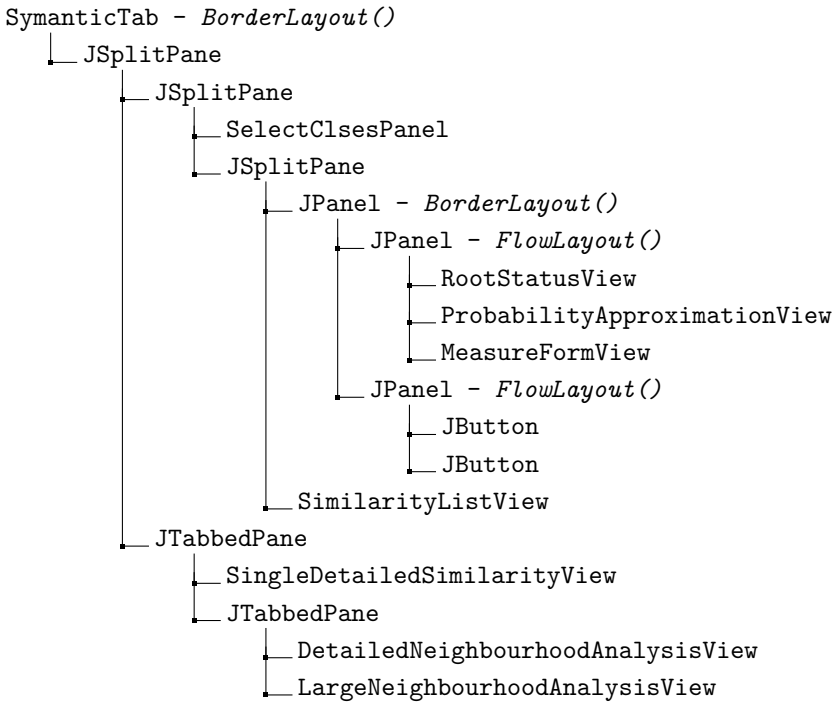


FIG. 8.22 – Arbre des widgets du plug-in SymanticTab

- le modèle est l'objet application ;
- la vue est la présentation du modèle à l'écran ;
- le controleur détermine la manière dont l'interface utilisateur réagit aux entrées utilisateur.

L'architecture MVC relève du *design patterns* dont l'objectif est de fournir un moyen efficace de réutiliser du code orienté objet pour différents projets et par différents programmeurs [Coo98]. En plus de la réutilisation des composants développés, ce découplage permet d'accroître la flexibilité des programmes.

Le découplage entre la vue et le modèle est mis en place par le biais d'un mécanisme de souscription/notification. Une vue doit garantir que son apparence reflète l'état du modèle. Pour cela, lorsque les données du modèle changent, le modèle le signale aux vues qui en dépendent. En réponse, chaque vue a la possibilité de se mettre à jour. Cette approche permet de lier de multiples vues à un seul modèle pour en fournir différentes présentations. Il est également envisageable de définir de nouvelles vues pour un modèle sans réécrire ce modèle, ce qui participe à la flexibilité du logiciel.

Cette interaction entre modèle et vue est décrite plus généralement par le patron Observerateur (*Observer pattern*) implémenté en Java par le biais de l'interface[9] *Observer* et de la classe *Observable*. Le modèle étend la classe *Observable* (*Model extends Observable*) et prévient lors d'un changement toutes les vues qui l'observent grâce aux méthodes *setChanged()* et *notifyObservers(Object arg)* dont il a hérité. Le modèle hérite également de la méthode *addObser-*

[9]La notion d'interface désigne une classe purement abstraite

ver(Observer view) qui permet à chaque vue de s'enregistrer auprès du modèle comme observateur de manière à être mis au courant lors d'une modification des données du modèle. Une vue dérive de l'interface Observer (*View implements Observer*) pour être identifiée comme tel et implémente la méthode *update(Observable model, Object arg)* qui est appelée par le modèle lors de la notification de changement.

Du côté du modèle, chaque vue est un observateur qui dispose à ce titre seulement d'une méthode *update* permettant de lui notifier un changement. En revanche, la vue peut librement accéder à l'ensemble des méthodes publiques du modèle pour l'interroger sur son état dans le but de se mettre à jour.

L'application scrupuleuse de l'architecture MVC entraîne une multiplication problématique du nombre de classes. Cela a pour première conséquence de ralentir l'interface utilisateur. Du point de vue du programmeur, cela augmente significativement la taille du code donc le risque d'erreur mais surtout il devient plus difficile à appréhender du fait de l'enchevêtrement des trois parties du code. Nous avons fait face à ce problème en utilisant le mécanisme Java des classes anonymes qui permet de rattacher le code du contrôleur à celui de la vue. Notre plug-in suit donc plus exactement une architecture Modèle/Vue-Contrôleur traduite dans ses grandes lignes par la figure 8.23.

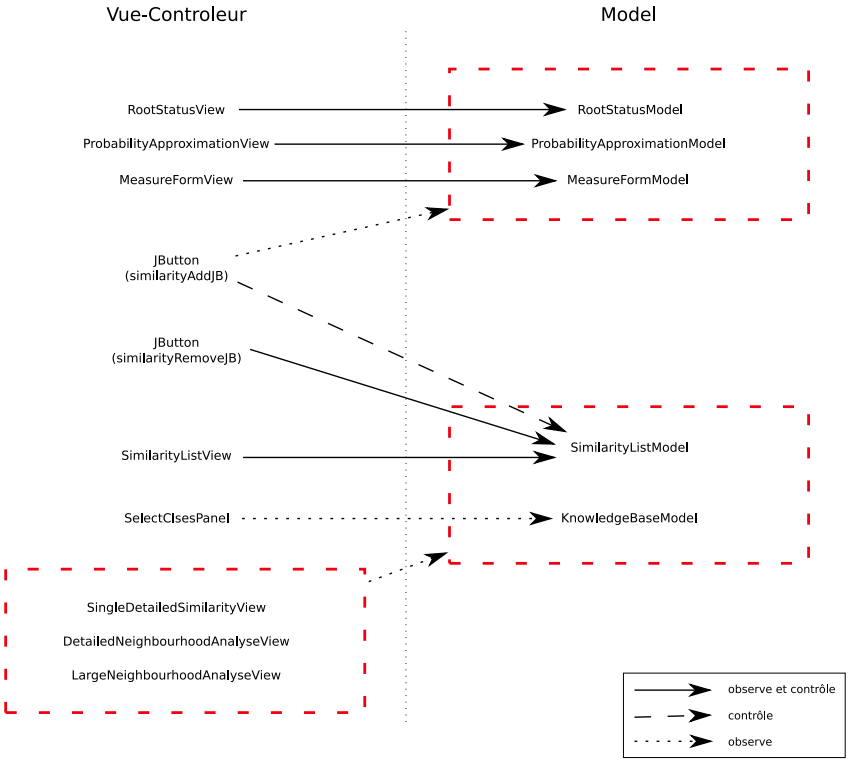


FIG. 8.23 – Architecture M/V-C du plug-in SymanticTab

8.5.5 Adaptation pour UEML

Le second plug-in baptisé *UEMLBase Correspondance Analyser* est donc une adaptation du plug-in SymanticTab qui a nécessité l'adaptation de l'ensemble des algorithmes mis en place pour le calcul des mesures, mais également un remaniement de l'interface. Au choix du statut de la racine, de l'approximation des probabilités et de la forme de la mesure s'ajoute la mise en place des coefficients qui fixent l'importance des concepts de chacune des hiérarchies UEML (cf. figure 8.24).

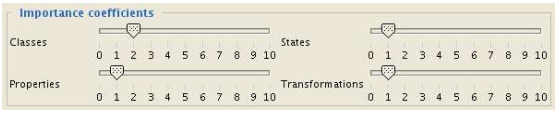
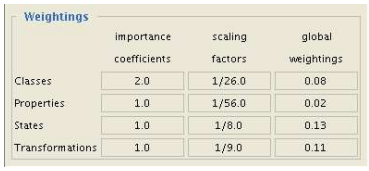


FIG. 8.24 – Paramétrage de l'importance des classes, propriétés, états et transformations

Lorsque l'utilisateur sélectionne une mesure préalablement définie, un panneau récapitule les pondérations qu'elle met en oeuvre (cf. figure 8.25). Il reprend les coefficients d'importance définis par l'utilisateur, les facteurs d'échelle et les pondérations globales ($\phi_c \cdot w_c$, $\phi_p \cdot w_p$, $\phi_e \cdot w_e$ et $\phi_t \cdot w_t$). Les vues destinées à l'analyse des mesures ont été légèrement modifiées pour fournir lorsque nécessaire le détail du calcul des contenus informationnel dans les quatre hiérarchies UEML.


Weightings

	importance coefficients	scaling factors	global weightings
Classes	2.0	1/26.0	0.08
Properties	1.0	1/56.0	0.02
States	1.0	1/8.0	0.13
Transformations	1.0	1/9.0	0.11

FIG. 8.25 – Visualisation des pondérations associées à une mesure donnée

8.6 Conclusion

Dans ce chapitre, nous avons présenté une application Java intégré à l'outil Protégé sous forme d'un plug-in de type *tabwidget* baptisée SymanticTab. Cette extension des fonctionnalités de Protégé permet l'analyse et la comparaison de similarités sémantiques sur une hiérarchie de subsomption donnée. Il s'agit d'un support pertinent pour guider le choix d'une similarité adaptée à la problématique de l'utilisateur.

Nous avons présenté l'interface homme-machine propre à notre plug-in et les grandes lignes de son implémentation. Nous avons notamment mis en avant la réutilisation de nombreux composants Java ainsi que le détail de son architecture M/V-C.

Nous avons également présenté une adaptation de notre plug-in pour une application en modélisation d'entreprise nécessitant l'analyse de mesures sémantiques (et non seulement des similarités) entre sous-ensemble de concepts. C'est d'ailleurs cette application développée dans le cadre du réseau d'excellence INTEROP-NoE qui a permis d'alimenter nos réflexions théoriques sur le sujet.

Conclusion

Les ontologies au coeur des systèmes à base de connaissances sont structurées par la relation de subsomption. Une mesure sémantique rend compte de la force d'une liaison entre deux concepts ou deux sous-ensembles de concepts de manière synthétique. Une telle évaluation de liaisons entre concepts permet de constituer une connaissance heuristique directement exploitable par des algorithmes adaptés. La mise en place de cette forme d'exploitation de la hiérarchie de subsomption se résume donc au choix d'une mesure appropriée pour atteindre les objectifs visés.

Contributions

Cette thèse participe à une meilleure compréhension des mesures sémantiques pour l'exploitation d'une hiérarchie de subsomption. Nous avons fourni un cadre formel pour la définition de mesures sémantiques et un outil pour l'analyse et la comparaison de ces mesures sur une hiérarchie réelle.

Au cours de ce manuscrit, nous avons traité des points suivants :

- Le contenu informationnel de Resnik [Res93] traduit la relation entre l'intension et l'extension d'un concept. Il s'agit d'un concept clef pour la définition des mesures sémantiques.
- L'exploitation d'un corpus (conjointement à la hiérarchie de subsomption) par le biais du contenu informationnel correspond à une approximation particulière de la probabilité attachée aux concepts. Nous envisageons d'autres approximations ne nécessitant pas de corpus.
- La quantité d'information extraite du corpus (en calculant le contenu informationnel à la manière de Resnik) est difficile à évaluer. De plus, la pertinence de cette information n'est pas une évidence.
- Nous avons approfondi la notion de contenu informationnel et les approximations proposées pour traiter correctement l'héritage multiple.
- Les mesures sémantiques de la littérature suivent des schémas classiques (e.g. coefficient de Jaccard, coefficient de Dice). Une analogie avec une représentation ensembliste met cela en lumière et contribue à préciser la signification des mesures sémantiques de la littérature.
- Nous avons adapté grâce à notre analogie les familles de similarité usuelles σ_α et σ_θ. Nous avons étudié leur comportement sur une hiérarchie de subsomption.
- L'analogie proposée offre également une voie d'investigation pour la défi-

nition de mesures sémantiques asymétriques. En effet, les travaux sur la qualité des règles en ECD peuvent potentiellement être adaptés dans cet esprit.
- Nous avons proposé un plug-in pour le logiciel Protégé qui reprend l'approche théorique développée dans cette thèse. L'interface utilisateur permet la définition d'une ou plusieurs similarités par le biais de divers paramètres dont le choix de l'approximation ou la forme de la mesure. Après avoir chargé une hiérarchie réelle sous Protégé, une seconde partie de l'interface offre différents angles d'analyse des mesures préalablement définies.
- Nous avons également proposé une adaptation de notre plug-in pour la base de connaissance UEML développée sous Protégé dans le cadre du réseau d'excellence INTEROP-NoE. Cette adaptation a nécessité l'implémentation d'une généralisation des mesures pour l'évaluation de liaisons entre deux sous-ensembles de concepts. Nous avons également étendu l'interface pour permettre la définition de mesures asymétriques.

Perspectives

Approfondissement des mesures asymétriques

Le problème de l'évaluation asymétrique dans une hiérarchie de subsomption est abordé par Rodriguez et Egenhofer [RE04]. L'analogie que nous avons proposé permet d'adapter un certain nombre d'indices de qualité de règles souvent asymétriques. On peut faire deux constats concernant les indices de règles asymétriques :

1. leur adaptation peut nous amener à des mesures parfois sans grand intérêt pour l'exploitation de la hiérarchie de subsomption ;
2. les variations dans la signification de ces indices sont beaucoup plus importantes qu'entre les mesures de ressemblance.

Nous avons cependant la ferme conviction que certains indices comme les indices descriptifs d'écart à l'équilibre peuvent fournir une fois adaptés des mesures sémantiques asymétriques avec des comportements tout à fait intéressants. La réutilisation de ces indices devrait permettre d'obtenir des indices avec une signification maîtrisée. Il reste toutefois que le comportement de ces indices mérite d'être étudié en détail.

Visualisation pour le développement d'ontologies

La définition d'une mesure sémantique selon l'approche défendue dans cette thèse nécessite le choix d'une approximation. En utilisant le contenu informationnel, chaque approximation permet de déterminer la spécificité des concepts et fournit ainsi une représentation de la hiérarchie. Une interface de visualisation peut être mise en place sous Protégé pour représenter une hiérarchie réelle selon une approximation donnée. On pourrait aussi proposer d'autres vues non pas basées sur le contenu informationnel mais sur le nombre de caractéristiques de chaque concept (en utilisant l'héritage).

De manière plus générale, une interface de développement d'une ontologie reposant sur une visualisation de la hiérarchie grâce à ces approximations est tout à fait envisageable. Cela nécessiterait de mettre en place des algorithmes de recalcul incrémental des approximations pour s'adapter aux modifications de la hiérarchie en temps réel. On peut également s'interroger sur l'utilisation de métaphores visuelles reposant sur une mesure sémantique (ou plusieurs mesures sémantiques complémentaires) pour naviguer dans l'ontologie.

Le développement d'ontologies est un processus lourd et la visualisation à base de mesures sémantiques nous semble un axe de recherche prometteur pour guider cette tâche.

Bibliographie

[ABH+08] V. Anaya, G. Berio, M. Harzallah, P. Heymans, R. Matulevicius, A. Opdahl, H. Panetto & M. Verdecho – « The unified enterprise modelling language – overview and further work », in *Proceedings of IFAC World Congress*, 2008, à paraître.

[AIS93] R. Agrawal, T. Imielienski & A. Swami – « Mining association rules between sets of items in large databases », in *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, ACM Press, 1993, p. 207–216.

[AN62] A. Arnault & P. Nicole – *La logique ou l'art de penser*, Flammarion, 1662.

[Aze03] J. Aze – « Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances », *Revue des Sciences et Technologies de l'Information* **17** (2003), no. 1-3, p. 171–182, Actes des journées Extraction et Gestion des Connaissances (EGC) 2003.

[BA99] R. J. Bayardo & R. Agrawal – « Mining the most interesting rules », in *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, ACM Press, 1999, p. 145–154.

[Bac04] B. Bachimont – « Pourquoi n'y a-t-il pas d'expérience en ingénierie des connaissances ? », in *Actes de la conférence Ingénierie des connaissances (IC'2004)*, Presses Universitaires de Grenoble, 2004.

[BB32] J. Braun-Blanquet – *Plant sociology : The study of plants communities*, New-York : McGraw-Hill, 1932.

[BGBG05] J. Blanchard, F. Guillet, H. Briand & R. Gras – « Assessing rule interestingness with a probabilistic measure of deviation from equilibrium », in *Proceedings of the eleventh international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, ENST, 2005, p. 191–200.

[BGGB04a] J. Blanchard, F. Guillet, R. Gras & H. Briand – « Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC », *Revue des Nouvelles Technologies de l'Information* **E-2** (2004), p. 287–298, Actes des journées Extraction et Gestion des Connaissances (EGC'2004).

[BGGB04b] — , « Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC », *Revue des Nouvelles Technologies de*

l'Information **E-2** (2004), p. 287–298, Actes des journées Extraction et Gestion des Connaissances (EGC) 2004.

[BH01] A. Budanitsky & G. Hirst – « Semantic distance in wordnet : An experimental, application-oriented evaluation of five measures », in *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics*, 2001.

[Bis00] G. Bisson – « La similarité : une notion symbolique/numérique. apprentissage symbolique-numérique (tome 2) », chapitre XX, Cépaduès, 2000.

[BKHB06] E. Blanchard, P. Kuntz, M. Harzallah & H. Briand – « A tree-based similarity for evaluating concept proximities in an ontology », in *Proc. of 10th conf. of the Int. Fed. of Classification Soc.*, Springer, 2006, p. 3–11.

[Bla05] J. Blanchard – « Un système de visualisation pour l'extraction, l'évaluation et l'exploration interactives de règles d'association », Thèse, École Polytechnique de l'Université de Nantes, 2005.

[BLHL01] T. Berners-Lee, J. Hendler & O. Lassila – « The semantic web », *Scientific American* (2001).

[BM51] R. Bush & F. Mosteller – « A model for stimulus generalization and discrimination », *Psychological Review* **58** (1951), p. 413–423.

[BM70] M. Barbut & B. Monjardet – *Ordre et classification – algèbre et combinatoire (2 tomes)*, Hachette, Paris, 1970.

[BM95] V. Batagelj & B. Matevž – « Comparing resemblance measures », *Journal of classification* **12** (1995), no. 1, p. 73–90.

[BMNPS91] F. Baader, D. Mcguinness, D. Nardi & P. Patel-Schneider – *The description logic handbook : Theory, implementation and applications*, Cambridge University Press, 1991.

[BMS97] S. Brin, R. Motwani & C. Silverstein – « Beyond market baskets : generalizing association rules to correlations », *SIGMOD Record* **26** (1997), no. 2, p. 265–276.

[BMUT97] S. Brin, R. Motwani, J. D. Ullman & S. Tsur – « Dynamic itemset counting and implication rules for market basket data », *SIGMOD Record* **26** (1997), no. 2, p. 255–264.

[BMV88] P. Bullen, D. S. Mitrinovic & P. M. Vasics – *Means and their inequalities*, Reidel Pub., 1988.

[BOAD04] G. Berio, A. Opdahl, V. Anaya & M. Dassisti – « UEML 2.0 », Deliverable 5.1 – INTEROP Network of Excellence, 2004, IST (Confidential).

[Bor97] W. N. Borst – « Construction of engineering ontologies », Thèse, University of Twente, 1997.

[Bou96] I. Bournaud – « Regroupement conceptuel pour l'organisation de connaissances », Thèse, Université Paris 6, 1996.

[BUB76] C. Baroni-Urbani & M. Buser – « Similarity of binary data », *Systematic Zoology* **25** (1976), p. 251–259.

[Bud99] A. Budanitsky – « Lexical semantic relatedness and its application in natural language processing », Tech. report, Computer Systems Research Group - University of Toronto, 1999.

[Bun79] M. Bunge – *Treatise on basic philosophy. ontology II : A world of systems*, Reidel, 1979.

[BYN99] R. Baeza-Yates & B. R. Neto – *Modern information retrieval*, addison-wesley éd., ACM Press Books, 1999.

[CDKFZ04] O. Corby, R. Dieng-Kuntz & C. Faron-Zucker – « Querying the semantic web with corese search engine », in *Proceedings of the 16th European Conference in Artificial Intelligence (ECAI'2004)*, 2004, p. 705–709.

[CFF+] V. Chaudhri, A. Farquhar, R. Fikes, P. Karp & J. Rice – « OKBC : a programmatic foundation for knowledge base interoperability », in *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI'98)*.

[CK96] F. Caillez & P. Kuntz – « A contribution to the study of the metric and euclidiean structures of dissimilarities », *Psychometrika* **61** (1996), no. 2, p. 241–253.

[CL75] A. Collins & E. Loftus – « A spreading activation theory of semantic processing », *Psychological Review* (1975), p. 407–428.

[Coh60] J. Cohen – « A coefficient of agreement for nominal scales », *Educational and Psychological Measurement* (1960), no. 20, p. 37–46.

[Con01] T. G. O. Consortium – « Creating the gene ontology resource : design and implementation », *Genome Res.* **11** (2001), p. 1425–1433.

[Coo98] J. W. Cooper – *The design patterns java companion*, Addison-Wesley Design Patterns Series, 1998.

[CQ69] A. M. Collins & M. R. Quillian – « Retrieval time from semantic memory », *Journal of Verbal Behavior and Verbal Learning* **8** (1969), p. 240–247.

[Dic45] L. R. Dice – « Measures of the amount of ecologic association between species », *Ecology* **26** (1945), no. 3, p. 297–302.

[EE59] H. Eisler & G. Ekman – « A mechanism of subjective similarity », *Acta Psychologica* **16** (1959), p. 1–10.

[eMM92] M. C. et M.L. Mugnier – « Conceptuals graphs : fundamental notions », *Revue d'Intelligence Artificielle (RIA)* **6** (1992), no. 4, p. 365–406.

[Fel98] C. Fellbaum (éd.) – *Wordnet : An electronic lexical database*, MIT Press, 1998.

[FGM+02] L. Finkelstein, E. Gabrilovich, Y. Matias, G. W. E. Rivlin, Z. Solan & E. Ruppin – « Placing search in context : The concept revisited », *ACM Transactions on Information Systems* **20** (2002), no. 1, p. 116–131.

[FGPJ97] M. Fernandez, A. Gomez-Perez & N. Juristo – « Methontology : from ontological art towards ontological engineering », in *Proceedings of the Spring Symposium Series on Ontological Engineering (AAAI'97)*, 1997.

[FJ65] S. Fillenbaum & L. V. Jones – « Grammatical contingencies in word association », *Journal of Verbal Learning and Verbal Behavior* **4** (1965), p. 248–255.

[Fre71] G. Frege – *Ecrits logiques et philosophiques*, 1971.

[Für04] F. Fürst – « Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation », Thèse, Ecole polytechnique de l'université de Nantes, 2004.

[Gan91] J.-G. Ganascia – « Charade : apprentissage de bases de connaissances », in *Induction symbolique et numérique à partir de données* (Y. Kodratoff & E. Diday, éds.), Cépaduès Editions, 1991, p. 309–326.

[GCB+04] R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz & P. Peter – « Quelques critères pour une mesure de qualité de règles d'association », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 3–31, numéro spécial Mesures de qualité pour la fouille de données.

[GCDKG05] F. Gandon, O. Corby, R. Dieng-Kuntz & A. Giboin – « Proximité conceptuelle et distances de graphes », in *Actes de l'atelier Raisonner le Web Sémantique avec des Graphes des 16e journées francophones d'Ingénierie des connaissances (IC'2005)*, 2005.

[GG95] N. Guarino & P. Giaretta – « Ontologies and knowledge bases : Towards a terminological clarification », in *Towards Very Large Knowledge Bases : Knowledge Building and Knowledge Sharing*, IOS Press, 1995, p. 25–32.

[GHJV95] E. Gamma, R. Helm, R. Johnson & J. M. Vlissides – *Design patterns : Elements of reusable object-oriented software*, Addison-Wesley Professional Computing Series, 1995.

[GL86] J. Gower & P. Legendre – « Metric and euclidean properties of dissimilarity coefficients », *Journal of Classification* **3** (1986), no. 1, p. 5–48.

[GN87] M. R. Genesereth & N. J. Nilsson – *Logical foundations of artificial intelligence*, Morgan Kaufmann, 1987.

[Gre75] R. Gregson – *Psychometrics of similarity*, Academic Press, 1975.

[Gro02] D. Grosser – « Construction itérative de bases de connaissances descriptives et classificatoires avec la plate-forme à objets IKBS - application à la systématique des coraux des mascareignes », Thèse, Université de La Réunion, 2002.

[Gru93] T. R. Gruber – « A translation approach to portable ontology specifications », *Knowledge Acquisition* **5** (1993), no. 2, p. 199–220.

[Gua98] N. Guarino – « Formal ontology in information systems », in *Proceedings of FOIS'98*, IOS Press, 1998, p. 3–15.

[Gui04] F. Guillet – *Mesures de la qualité des connaissances en ECD*, 2004, Tutoriel des journées Extraction et Gestion des Connaissances, www.isima.fr/~egc2004/Cours/Tutoriel-EGC2004.pdf.

[GW99] B. Ganter & R. Wille – *Formal concept analysis*, Springer, Berlin, 1999.

[Ham61] U. Hamann – « Merkmalsbestand und verwandtschaftsbeziehungen der farinosae. ein beitrag zum system der monokotyledonen », *Willdenowia* **2** (1961), p. 639–768.

[HSO98] G. Hirst & D. St-Onge – « Lexical chains as representation of context for the detection and correction of malaproprisms », in *WordNet : An electronic lexical database* (C. Fellbaum, éd.), MIT Press, 1998, p. 305–332.

[Jac01] P. Jaccard – « Distribution of the alpine flora in the dranse's basin and some neighbouring regions (in french) », *Bulletin de la Soc. Vaudoise Sci. Nat.* (1901), no. 37, p. 241–272.

[JC97] J. J. Jiang & D. W. Conrath – « Semantic similarity based on corpus statistics and lexical taxonomy », in *Proc. of int. conf. on Research in Computational Linguistics*, 1997, p. 19–33.

[KK90] Y. Kim & J. Kim – « A model of knowledge based information retrieval with hierarchical concept graph », *Journal of Documentation* **46** (1990), no. 2, p. 113–136.

[KLW95] M. Kifer, G. Lausen & J. Wu – « Logical foundations of object-oriented and frame-based languages », *Journal of the ACM* **42** (1995), no. 4, p. 741–843.

[KP88] G. E. Krasner & S. T. Pope – « A cookbook for using the model-view controller user interface paradigm in smalltalk-80 », *Journal of Object-Oriented Programming* **1** (1988), no. 3.

[Kul28] S. Kulczynski – « Zespoly roslin w pieninach–die pflanzenassoziationen der pienenen », *Bulletin international de l'académie polonaise des sciences et des lettres, Classe des Sciences Mathématiques et Naturelles, série B, Supplément II* (1928), p. 57–203.

[LBM03] Y. Li, Z. A. Bandar & D. McLean – « An approach for measuring semantic similarity between words using multiple information sources », *IEEE Trans. on Knowledge and data engineering* **15** (2003), no. 4, p. 871–882.

[LC94] C. Leacock & M. Chodorow – « Filling in a sparse training space for word sense identification », march 1994.

[LC98] — , « Combining local context and wordnet similarity for word sense identification », in *WordNet : An electronic lexical database* (C. Fellbaum, éd.), MIT Press, 1998, p. 265–283.

[LFZ99] N. Lavrac, P. A. Flach & B. Zupan – « Rule evaluation measures : a unifying view », in *ILP'99 : Proceedings of the ninth International Workshop on Inductive Logic Programming*, Springer-Verlag, 1999, p. 174–185.

[LH05] M. Laukkanen & H. Helin – « Competence management within and between organizations », in *Proc. of 2nd interop-EMOI*

Workshop on Enterprise Models and Ontologies for Interoperability at the 17th conf. on advanced information systems engineering, Springer, 2005, p. 359–362.

[Lin98] D. Lin – « An information-theoretic definition of similarity », in *Proc. of the 15th int. conf. on machine learning*, Morgan Kaufmann, 1998, p. 296–304.

[LKL93] J. H. Lee, M. H. Kim & Y. J. Lee – « Information retrieval based on conceptual distance in is-a hierarchies », *Journal of Documentation* **49** (1993), no. 2, p. 188–207.

[Loe47] J. Loevinger – « A systematic approach to the construction and evaluation of tests of ability », *Psychological Monographs* **61** (1947), no. 4.

[LSBG03] P. Lord, R. Stevens, A. Brass & C. Goble – « Investigating semantic similarity measures across the gene ontology : the relationship between sequence and annotation », *Bioinformatics* **19** (2003), no. 10, p. 1275–1283.

[LT04] S. Lallich & O. Teytaud – « Evaluation et validation de l'intérêt des règles d'association », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 193–218, numéro spécial Mesures de qualité pour la fouille de données.

[MBF+90] G. Miller, R. Beckwith, C. Fellbaum, D. Gross & K. Miller – « Introduction to wordnet : An on-line lexical database », *International Journal of Lexicography* **3** (1990), p. 235–312.

[MC91] G. Miller & W. Charles – « Contextual correlates of semantic similarity », *Language and Cognitive Processes* **6** (1991), no. 1, p. 1–28.

[MH91] J. Morris & G. Hirst – « Lexical cohesion computed by structure of text », *Computational Linguistics* **17** (1991), no. 1, p. 21–48.

[Mic20] E. L. Michael – « Marine ecology and the coefficient of association : A plea in behalf of quantitative biology », *The Journal of Ecology* **8** (1920), no. 1, p. 54–59.

[Mil85] G. A. Miller – « Wordnet : A dictionary browser in information in data », in *Proceedings of the First Conference of the UW Centre for the New Oxford Dictionary*, 1985.

[Min75] M. Minsky – *A framework for representing knowledge*, The psychologie of Computer Vision, 1975.

[MMRV05] A. G. Maguitman, F. Menczer, H. Roinestad & A. Vespignani – « Algorithmic detection of semantic similarity », in *Proc. of the 14th int. conf. on world wide web*, ACM Press, 2005, p. 107–116.

[MTR89] C. F. McMath, R. S. Tamaru & R. Rada – « A graphical thesaurus-based information retrieval system », *International Journal of Man-Machine Studies* **31** (1989), no. 2, p. 121–147.

[Nap97] A. Napoli – « Une introduction aux logiques de descriptions », Tech. Report 3314, INRIA, 1997.

[Neb90] B. Nebel – « Reasoning and revision in hybrid representation systems », in *Lecture Notes in Artificial Intelligence*, vol. 422, Springer-Verlag, 1990.

[NFM00] N. F. Noy, R. W. Fergerson & M. A. Musen – « The knowledge model of protege-2000 : Combining interoperability and flexibility », in *2th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, 2000.

[OB06] A. Opdahl & G. Berio – « Interoperable language and model management using the UEML approach », in *Proceedings of Workshop on Global Integrated Model Management*, 2006.

[Och57] A. Ochiaï – « Zoogeographic studies of the soleoid fishes found in japan and its neighbouring regions », *Bulletin of the Japanese Society for Scientific Fisheries* **22** (1957), p. 526–530.

[PD02] M. Petit & G. Doumeingts – « Report on the state of the art in enterprise modelling », Tech. report, University of Namur, 2002.

[Pea96] K. Pearson – « Mathematical contributions to the theory of evolution : regression, heredity and panmixia », *Philosophical Transactions of the Royal Society Of London* **series A** (1896), no. 187, p. 253–318.

[PPM04] T. Pedersen, S. Patwardhan & J. Michelizzi – « Wordnet : :similarity - measuring the relatedness of concepts », in *In proc. of the Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 2004, p. 38–41.

[Qui68] M. Quillian – « Semantic memory », in *in Semantic Information Processing*, 1968.

[RE04] A. Rodríguez & M. Egenhofer – « Comparing geospatial entity classes : An asymmetric and context-dependent similarity measure », *International Journal of Geographical Information Science* **18** (2004), no. 3, p. 229–256.

[Res61] F. Restle – *Psychology of judgment and choice*, New York : Wiley, 1961.

[Res93] P. Resnik – « Selection and information : A class based approach to lexical relationships », Thèse, University of Pennsylvania, 1993.

[Res95] — , « Using information content to evaluate semantic similarity in a taxonomy », in *Proc. of the 14th int. Joint conf. on Artificial Intelligence*, vol. 1, 1995, p. 448–453.

[Res99] — , « Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language », *Journal of Artificial Intelligence Research* **11** (1999), p. 95–130.

[RG65] H. Rubenstein & J. Goodenough – « Contextual correlates of synonymy », *Communications of the ACM* **8** (1965), no. 10, p. 627–633.

[RMBB89] R. Rada, H. Mili, E. Bicknell & M. Blettner – « Development and application of a metric on semantic nets », *IEEE Transactions on Systems, Man, and Cybernetics* **19** (1989), no. 1, p. 17–30.

[RR40] P. Russel & T. Rao – « On habitat and association of species of anopheline larvae in south-eastern madras », *Journal of the Malaria Institute of India* (1940), no. 3, p. 153–178.

[RS95] R. Richardson & A. F. Smeaton – « Using wordnet in a knowledge-based approach to information retrieval », Tech. Report CA-0395, School of Computer Applications, Dublin, Ireland, 1995.

[RT60] D. Rogers & T. Tanimoto – « A computer program for classifying plants », *Science* (1960), no. 132, p. 1115–1118.

[SBF98] R. Studer, V. R. Benjamins & D. Fensel – « Knowledge engineering : Principles and methods », *Data and Knowledge Engineering* **25** (1998), p. 161–197.

[SBT+06] O. Steichen, C. D.-L. Bozec, M. Thieu, E. Zapletal & M.-C. Jaulent – « Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus », *Computers in Biology and Medicine* **36** (2006), no. 7-8, p. 768–788.

[Sim60] G. G. Simpson – « Notes on the measurement of faunal resemblance », *American Journal of Science* **258-A** (1960), p. 300–311.

[Sjö72] L. Sjöberg – « A cognitive theory of similarity », *Göteborg Psychological Reports* **2** (1972), no. 10.

[SM58] R. Sokal & C. Michener – « A statistical method for evaluating systematic relationships », *University of Kansas Science Bulletin* (1958), no. 38, p. 1409–1438.

[Smi78] E. E. Smith – « Theories of semantic memory », In Estes, W. K. (ed.). Handbook of Learning and Cognitive Processes, vol. 5. Hillsdale, NJ : Erlbaum. The Synonym Finder. 1978. Emmaus, Pa. : Rodale Press., 1978.

[SMS+01] N. Stojanovic, A. Maedche, S. Staab, R. Studer & Y. Sure – « Seal : a framework for developing semantic portals », in *Proc. of the int. conf. on Knowledge capture*, 2001, p. 155–162.

[Sow84] J. Sowa – *Conceptual structures : Information processing in mind and machine*, Addison-Wesley, 1984.

[SS63] R. R. Sokal & P. H. Sneath – *Principles of numerical taxonomy*, W. H. Freeman and Co, 1963.

[SS88] M. Sebag & M. Schoenauer – « Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases », in *Proceedings of the European knowledge acquisition workshop EKAW'88*, Gesellschaft für Mathematik und Datenverarbeitung mbH, 1988, p. 28.1–28.20.

[Sus93] M. Sussna – « Word sense disambiguation for free-text indexing using a massive semantic network », in *Proc. of the Second International Conference on Information and Knowledge Management*, 1993, poster, p. 67–74.

[SVH04] N. Seco, T. Veale & J. Hayes – « An intrinsic information content metric for semantic similarity in wordnet », in *Proc. of the 16th european conf. on artificial intelligence*, 2004, p. 1089–1090.

[SW49] C. Shannon & W. Weaver – *The mathematical theory of communication*, University of Illinois Press, 1949.

[TSZ+04] M. Thieu, O. Steichen, E. Zapletal, M.-C. Jaulent & C. L. Bozec – « Mesures de similarité pour l'aide au consensus en anatomie pathologique », in *15es journées francophones d'Ingénierie des Connaissances (IC'2004)*, Presses Universitaires de Grenoble, 2004, p. 225–236.

[Tve77] A. Tversky – « Features of similarity », *Psychological Review* **84** (1977), no. 4, p. 327–352.

[Ver02] F. Vernadat – « UEML : towards a unified enterprise modelling language », *International Journal of Production Research* **40** (2002), no. 17, p. 4309–4321.

[WG01] C. Welty & N. Guarino – « Supporting ontological analysis of taxonomic relationships », *Data and Knowledge Engineering* **39** (2001), no. 1, p. 51–74.

[Wil82] R. Wille – « Restructuring lattice theory : an approach based on hierarchies of concepts, in ordered sets », p. 445–470, Reidel, 1982.

[WP94] Z. Wu & M. Palmer – « Verb semantics and lexical selection », in *Proc. of the 32nd annual meeting of the associations for Comp. Linguistics*, 1994, p. 133–138.

[WW93] Y. Wand & R. Weber – « On the ontological expressiveness of information systems analysis and design grammars », *Journal of Information Systems* (1993).

[Yul00] G. Yule – « On the association of attributes in statistics », *Philosophical Transactions of the Royal Society of London* **series A** (1900), no. 194, p. 257–319.

[ZZLY02] J. Zhong, H. Zhu, J. Li & Y. Yu – « Conceptual graph matching for semantic search », in *Proceedings of the 10th International Conference on Conceptual Structures (ICCS'02)* (London, UK), Springer-Verlag, 2002, p. 92–106.

Résumé :
De nombreux travaux en Ingénierie des Connaissances reposent sur le développement puis l'exploitation d'ontologies. Les connaissances qu'elles renferment sont structurées autour de la relation de subsomption qui définit une structure hiérarchique. Cette hiérarchie de subsomption est parfois exploitée à l'aide d'une mesure sémantique qui fournit une évaluation numérique d'une liaison entre deux concepts ou deux sous-ensembles de concepts.
On trouve dans la littérature diverses mesures généralement définies de manière *ad hoc* pour les besoins d'une application spécifique. La diversité des domaines considérés rend complexe la comparaison des mesures existantes ainsi que leur réutilisation. Il est difficile de faire un choix avisée sur la base des travaux existants sans le recul nécessaire. Cependant, le choix préalable d'une « bonne » mesure est un problème important puisqu'il influe sur la pertinence des résultats obtenus en aval.
Notre thèse pose un cadre théorique qui supporte l'analyse, la comparaison et la définition de mesures sémantiques. La singularité de notre approche est qu'elle repose sur l'utilisation du contenu informationnel sans toutefois nécessité l'utilisation d'un corpus. Nous proposons des approximations de la mesure de probabilité qui permettent d'exploiter divers aspects d'un arbre de subsomption. Nous approfondissons la notion de contenu informationnel et les approximations proposées pour traiter correctement le cas de l'héritage multiple.
Le problème de l'évaluation de liaisons entre deux objets sur la base d'une représentation ensembliste est largement traité dans la littérature. L'intérêt majeur de notre approche est de s'appuyer sur ces travaux grâce à une analogie qui permet de les transposer à une hiérarchie de subsomption. On met d'ailleurs en évidence des schémas classiques (e.g. coefficient de jaccard, coefficient de Dice) que respectent certaines mesures sémantiques. Cette analogie ouvre également la voie pour la définition de mesures sémantiques asymétriques en adaptant des travaux sur la qualité des règles en ECD (Extraction de Connaissances à partir des Données).
Nous présentons un outil sous forme d'un plug-in pour le logiciel Protégé qui reprend l'approche théorique développée dans cette thèse. Nous avons développé une interface utilisateur qui permet la définition d'une ou plusieurs similarités sémantiques par le biais de divers paramètres dont le choix de l'approximation et la forme de la mesure. Notre outil offre différents angles d'analyse du comportement de ces similarités sur une hiérarchie réelle préalablement chargée sous Protégé. Nous avons également proposé une adaptation de notre plug-in pour la base de connaissance UEML développée dans le cadre du réseau d'excellence INTEROP-NoE. Cette adaptation a nécessité l'implémentation d'une généralisation des mesures pour l'évaluation de liaisons entre deux sous-ensembles de concepts. Nous avons également étendu l'interface pour permettre la définition de mesures asymétriques.

Mots-clés :
Ingénierie des Connaissances,hiérarchie de subsomption,ontologie,mesures de ressemblance,mesures sémantiques

Abstract :
A lot of work in knowledge engineering is based on the ontology development and exploitation. In an ontology, the knowledge is organized around the subsumption relationship which describe a hierarchical structure. This subsumption hierarchy is sometimes exploited by the way of semantic measures which gave us a numerical evaluation of a link between two concepts or two concept subsets.
We find in the literature various measures usually defined in an *ad hoc* way to fullfil the need of a specific application. The diversity of considered domains make it difficult to compare the existing measures and to reuse it. It is difficult to make an acute choice without having the necessar hindsight on existing work. However, the preliminary choice of a " good " measure is a crucial problem because it influence the relevance of the downward obtained results.
In our thesis, we define a theoretical framework which support the analysis, the comparison and the definition of the information content. The uniqueness of our approach is that it exploits the information content without the necessity of a corpus. We propose several approximations of the probability measure which allow us to exploit various aspects of a subsumption tree. To deal properly with the multiple inheritance, we are going into detail of the information content and of the previously proposed approximations.
The problem of the evaluation of links between two objects represented by sets is widely discussed in the litterature. The major interest of our approach is to reuse this work by the way of an analogy which allows us to adapt it to a subsumption hierarchy. We underline usual form (e.g. Jaccard's coefficient, Dice's coefficient) which are followed by existing semantic measures. This analogy open the way to the definition of asymmetrical semantic measures by adapting so much work on quality measures for association rules in KDD (Knowledge Discovery in Databases).
We present a plug-in for the Protégé tool which taken back the theoretical approach developed in this thesis. We have implemented a user interface which allow us to define semantic similarities by the way of various parameters like the approximation choice or the measure form choice. Our tool offer different ways to analyse the behaviour of these similarities on a subsumption hierarchy beforehand loading under Protégé. We have also proposed an adaptation of our plug-in for the UEML knowledge base developed in conjunction with the INTEROP-NoE network of excellence. This adaptation need to implement a generalization of the measures to evaluate links between two concept subsets. We have also extended the interface to permit the definition of asymmetrical measures.

Keywords :
Knowledge engineering,subsumption hierarchy,ontology,resemblance measures,semantic measures